Volume 43 *January 1957* *Number 1*

Proceedings of the
National Academy of Sciences
of the United States of America

★ ★ ★ ★ ★

★ ★ ★ ★ ★

The Proceedings of the National Academy of Sciences

OFFICERS OF THE ACADEMY

DETLEV W. BRONK
President

GEORGE W. CORNER
Vice President

JOHN G. KIRKWOOD
Foreign Secretary

HUGH L. DRYDEN
Home Secretary

WILLIAM J. ROBBINS
Treasurer

EDITORIAL BOARD OF THE PROCEEDINGS

WENDELL M. STANLEY
Chairman

HUGH L. DRYEEN, *Home Secretary*
WILLIAM J. ROBBINS, *Treasurer*

JOHN G. KIRKWOOD, *Foreign Secretary*
EDWIN B. WILSON, *Managing Editor*

J. W. BEAMS	G. W. BEADLE	GREGORY BREIT
R. E. CLELAND	E. A. DOISY	L. C. DUNN
W. W. RUBEY	G. C. EVANS	W. F. LIBBY
HARLOW SHAPLEY	C. H. GRAHAM	J. VON NEUMANN
F. E. TERMAN	J. T. PATTERSON	PAUL WEISS

Subscription price is $12.50 per volume (12 issues); the price of a single issue is $1.25. Make all remittances payable to The University of Chicago Press in U.S. currency or its equivalent. Authorized agent for the British Commonwealth, except North America and Australasia: Cambridge University Press, Bentley House, 200 Euston Road, London, N. W. 1, England.

Editorial correspondence should be addressed to the PROCEEDINGS of the NATIONAL ACADEMY OF SCIENCES, Attention Miss Mary D. Alexander, University of Chicago Press, 5750 Ellis Avenue, Chicago 37, Illinois.

Business correspondence should be addressed to the Journals Division, University of Chicago Press, 5750 Ellis Avenue, Chicago 37, Illinois.

Subscribers are requested to notify The University of Chicago Press and their local postmaster immediately of change of address.

Microfilms of complete volumes of this journal are available to regular subscribers only and may be obtained at the end of the volume year from University Microfilms, 313 N. First Street, Ann Arbor, Michigan.

Entered as second-class matter June 21, 1920 at the Post-office at Easton, Pennsylvania, under the Act of August 24, 1912. Accepted for mailing at a special rate of postage provided for in Section 1103, Act of October 3, 1917. Authorized June 18, 1920.

PRINTED IN THE U. S. A.

THE PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES
is published for
THE NATIONAL ACADEMY OF SCIENCES
by
THE UNIVERSITY OF CHICAGO PRESS

Proceedings of the
NATIONAL ACADEMY OF SCIENCES

Volume 43 · Number 1 · January 15, 1957

GEOPHYSICS SYMPOSIUM ON SOLAR AND TERRESTRIAL RELATIONSHIPS

BY INVITATION OF THE COMMITTEE ON ARRANGEMENTS FOR THE ANNUAL MEETING

Read before the Academy, April 23, 1956

CONTENTS

INTRODUCTORY REMARKS	*Ernest H. Vestine*	1
INFERENCES FROM RADIO SIGNALS FROM THE SUN AND PLANETS	*John Firor*	2
THE SOLAR HYDROMAGNETIC DYNAMO	*E. N. Parker*	8
THE TERRESTRIAL DYNAMO	*Walter M. Elsasser*	14
ON COSMIC MAGNETIC FIELDS	*S. Chandrasekhar*	24
SOLAR INFLUENCES ON COSMIC RAYS	*S. E. Forbush*	28
SOLAR ORIGIN OF CHANGES IN THE PRIMARY COSMIC RADIATION	*J. A. Simpson*	42
DIRECT DETECTION OF AURORAL RADIATION WITH ROCKET EQUIPMENT	*James A. Van Allen*	57
PROBLEMS IN AURORAL MORPHOLOGY	*C. T. Elvey*	63
SOLAR INFLUENCES ON GEOMAGNETISM	*Julius Bartels*	75
OBSERVATIONAL AND THEORETICAL ASPECTS OF MAGNETIC AND IONOSPHERIC STORMS	*Ernest H. Vestine*	81
VARIATIONS OF THE IONOSPHERE AND OF THE SOLAR INTENSITY WITH SUNSPOTS	*E. O. Hulburt*	92
SOLAR RADIATION AND THE LOWER ATMOSPHERE	*Sigmund Fritz*	95

INTRODUCTORY REMARKS

BY ERNEST H. VESTINE

CARNEGIE INSTITUTION OF WASHINGTON

Symposia have been held since the time of ancient Greece. In those days a symposium implied a drinking together by a social gathering at which there was a free interchange of ideas on a particular subject. One of the dialogues of Plato reported such a symposium on ideal love. Today we are gathered together to discuss our love of the earth, and especially of a few facets of understanding we are achieving, linking events on and above the earth with others observable on the sun.

The subject of solar and terrestrial relationships in geophysics is a very broad one, since so many terrestrial manifestations of energy are closely linked with things solar. In many areas it is difficult though probably irrelevant to determine where astronomy ends and geophysics begins. In particular, there are numerous electromagnetic phenomena in common, and changes in the earth's atmosphere and ionosphere are profoundly influenced by solar radiation. Since much of human progress consists of ascertaining and understanding our environment, there is much to interest both the astronomer and the geophysicist. Moreover, the region above the earth's surface is becoming an increasingly accessible part of our environment. An era of man-made satellites is beginning which will provide new information supplementing older results based on radiation and rockets. It is hence a region in which rapid progress is likely to be close at hand.

The opportunities for study in this area of geophysics are considerable. This creates a need for people trained by universities to work in this area. These scientific investigators will be mainly physicists, engineers, astronomers, and mathematicians. They should also have courses and graduate theses in geophysics, now practically nonexistent in some of the areas to be discussed by specialists and others here today.

Today's symposium was suggested by Dr. John Simpson, of the University of Chicago, and other scientists actively engaged in these areas of research. The conference arrangements were made by Dr. Simpson and the writer at the request of Dr. L. V. Berkner as chairman of the Geophysics Section of the National Academy of Sciences. The program begins with the description of radio signals from the sun and planets, and what we learn from them. There follows a discussion of the subtle enigmas of solar magnetism and geomagnetism, and of the related influences of hydromagnetism upon both. These fields and influences profoundly affect cosmic rays, and inferences from these are noted. New results respecting low-energy radiation detected by means of rocket techniques are made known, as well as changes in aurora, geomagnetism, and the ionosphere which this radiation may produce. Finally, the very important aspect of the influence of the sun upon the weather at ground level is discussed in some of its most recent aspects.

INFERENCES FROM RADIO SIGNALS FROM THE SUN AND PLANETS

By John Firor

CARNEGIE INSTITUTION OF WASHINGTON, DEPARTMENT OF TERRESTRIAL MAGNETISM, WASHINGTON, D.C.

Introduction—The detailed nature of the solar corona is important in discussions of solar and terrestrial relationships, since any agency passing from sun to earth must pass through the corona and produce effects there, and some of these effects may be measurable. Furthermore, the suggestion has been made, by Professor Chapman, that the corona itself is of sufficient extent to produce effects directly on the earth's ionosphere.

Radio Brightness Distribution—Observations of the solar corona by optical means at the times of total eclipses lead to a picture of the corona as a completely ionized gas (mostly hydrogen) at a temperature of around a million degrees and a density which falls off roughly as the inverse sixth power of the distance from the center of the sun. The corona is not completely symmetrical but contains streamers of higher than average density. Also, at times of minimum solar activity, the corona is less dense near the poles of the sun than at the equator, and there is some evidence that the density varies by about a factor of 2 during the sunspot cycle. The optical eclipse observations give quantitative results in a rather limited range of distance from the sun—up to about 4 solar radii. Farther from the sun the true coronal light is obscured by light scattered by dust particles—the so-called zodiacal light.

One can also observe the corona in the radio-wave-length region. In fact, at meter wave lengths, the corona is all of the sun that one can observe, and no eclipse is required.

If the corona is actually constituted as the results from the optical eclipse observations indicate, then one may predict that the corona should be a source of radio energy, and its amount and distribution in angle as seen from the earth may be calculated.[1] These predictions are summarized in Figure 1. For very short wave

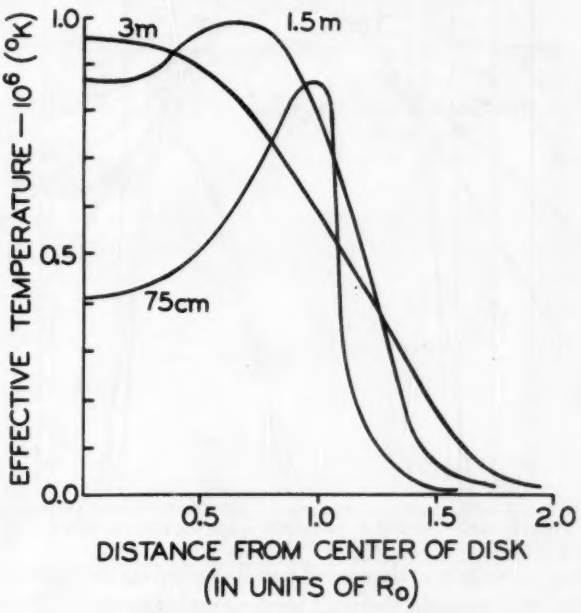


FIG. 1.—The computed brightness distribution of the sun at three wave lengths.

lengths these curves should reduce to the optical case in which the central disk temperature is around 5,000° and the temperature decreases slowly to the limb and then drops sharply.

The calculation of brightness distribution involves two steps. The first is to determine the ray trajectories. In the case of meter-wave-length radio waves, the variation of electron density, and hence of index of refraction, with distance from the sun is sufficient to produce considerable refraction. Thus the radiation we see coming from a particular direction does not all originate along a straight line in that direction. Figure 2 shows some ray paths calculated for waves of 16.5 meters wave length.

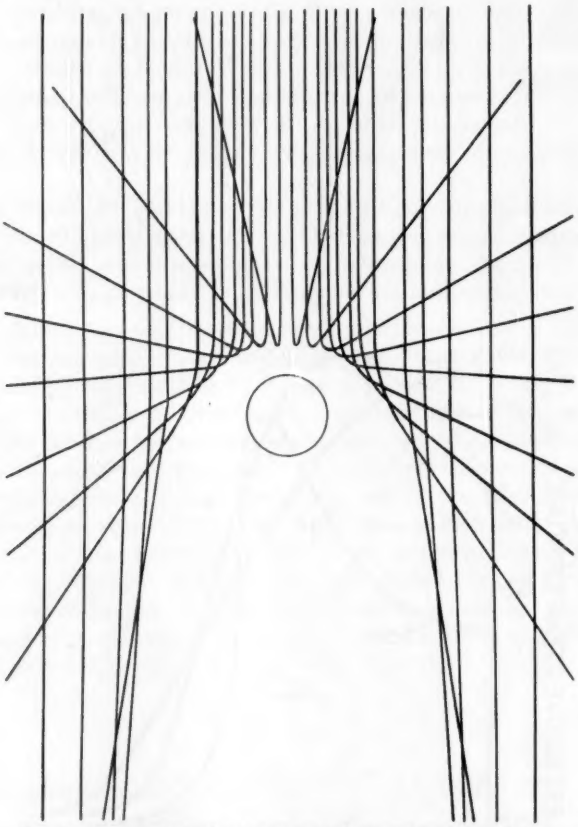


FIG. 2.—Ray paths in the corona at 16.5 meters wave length from a paper by Bracewell and Preston (*Astrophys. J.*, **123**, 14, 1956).

With the ray paths known, the second step is to compute the emission and absorption along the ray to obtain the brightness in that direction. The experimental problem is then to measure the brightness distribution and confirm or cast doubt on this calculation and its data and assumptions.

This procedure may seem backwards, since it would seem better to measure the brightness distribution at several wave lengths, deduce the electron density and temperature, and then compare with similar quantities deduced from optical observations. However, the strong refraction, or bending of the rays, and the depend-

ence of the emission on temperature as well as on electron density seem to make the solution of the necessary integral equation in the radio case impossible.

At meter wave lengths the measurement of the brightness distribution of the sun presents a problem in antenna size required. An antenna several hundred wave lengths in size is required. Hence these measurements have so far been made by a method similar to that used by Michelson for measuring stellar diameters. Two spaced aerials have a reception pattern which varies with angle θ as cos $(2\pi d/\lambda)\theta$, where d/λ is the spacing in wave lengths. Thus the output of a receiver attached to these two aerials is proportional to $\int B(\theta) \cos (2\pi d/\lambda)\theta \, d\theta$, where $B(\theta)$ is the brightness distribution in the sky. (The one-dimensional case is taken for simplicity.) The receiver output is, then, one Fourier component of the brightness distribution. By varying d/λ, all components can be measured and $B(\theta)$ reconstructed.

Measurements have been made by this method at meter wave lengths[2-4] and at shorter wave lengths by actually building an antenna hundreds of wave lengths in size.[5-7] The results are in good agreement with the calculations, with the exception that the radio sun is always somewhat larger than expected. Figure 3 compares the measured and calculated distributions at 1.5 meters wave length.[3] In general, the thermal radiation from the corona seems well understood.

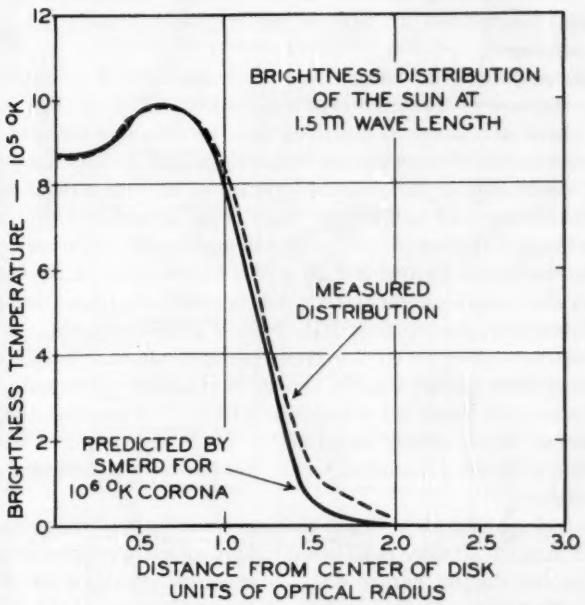


FIG. 3.—Brightness distribution of the sun as calculated by Smerd (*Australian J. Sci. Research*, *A*, **3**, 34, 1950) and as measured by the Fourier synthesis technique (Firor, *Astrophys. J.*, **123**, 320, 1956).

Occultation Of Radio Stars—The radio brightness distribution method explores roughly the same region of the corona as the optical observation. Radio methods

have also been used to observe regions of the corona farther from the sun.[8, 9] One can, using the same ray trajectories mentioned before, calculate the appearance of a radio star as seen through the solar corona. At the longer wave lengths it is found that absorption would be small and the source should be observable with only refractive effects present. We expect as the star nears the sun an image of the star to appear in the corona and the star itself to be magnified and shifted in position.[10] Finally the star should disappear behind the corona. The effect should depend strongly on wave length.

One of the strong radio sources, the Crab Nebula, passes within about a degree of the sun each year. Observations were made in England,[8,11] in Washington, and elsewhere to look for these four features of the predicted occultation: increase in intensity, shift in position, disappearance on the longer wave lengths (around 8 meters), and no effect on short wave lengths (for example, 3.7 meters). Not one of these predictions was fulfilled. The most striking feature of the observations was that the source began to be affected some 5° from the sun, several times farther away than predicted. This effect was an increase in angular size with no change in the total radiation. The increase in angular size also occurred on 3.7 meters where no effect should have been observed. On the longer wave lengths the source was still present when it should have been occulted, and no shift in position was found. In Washington the results were further complicated by an occasion when the source disappeared but returned suddenly to full intensity for an hour and then disappeared again.

All these things point to deviations from the simple, smooth corona which proved successful in the case of the brightness distribution. Hewish[9] has accounted for many of the observed features by assuming that the corona is irregular in structure, and he points to the visible streamers as evidence for such density fluctuations. The Washington observation of the temporary return of the source requires that the irregularities be so large and so few that there is an appreciable probability of *no* irregularities being in the line of sight. One should note that the part of the corona studied by this method is the region 4–20 radii from the center of the sun.

The corona also occults another strong radio source—the planet Jupiter. In this case the radiation from the source is in the form of pulses or bursts. In 1951 bursts from the planet were observed on two days when it should have been occulted. But here the most striking feature was the change in character of the radiation. In all the observations made when the source was within 5° of the sun, the bursts were smoothed into an almost steady radiation.[12] To do this requires relative delays in the radiation of about a minute of time. No one has yet proposed a method of getting such delays.

One can safely conclude from these two radio occultation observations that the corona in the range 4–20 solar radii is not simply an extrapolation of the optically derived corona, but the true nature of the outer corona remains a problem.

Active Solar Radiation—Perhaps the most complex feature of the radio corona is the active radiation originating there. The total radiation from the sun never drops below the level of the thermal radiation described earlier, but upon occasion it increases by a factor of a thousand or a million in a few seconds, as shown in Figure 4. These large increases of intensity are frequently associated with optically observed phenomena on the sun. particularly solar flares and with the attendant terrestrial

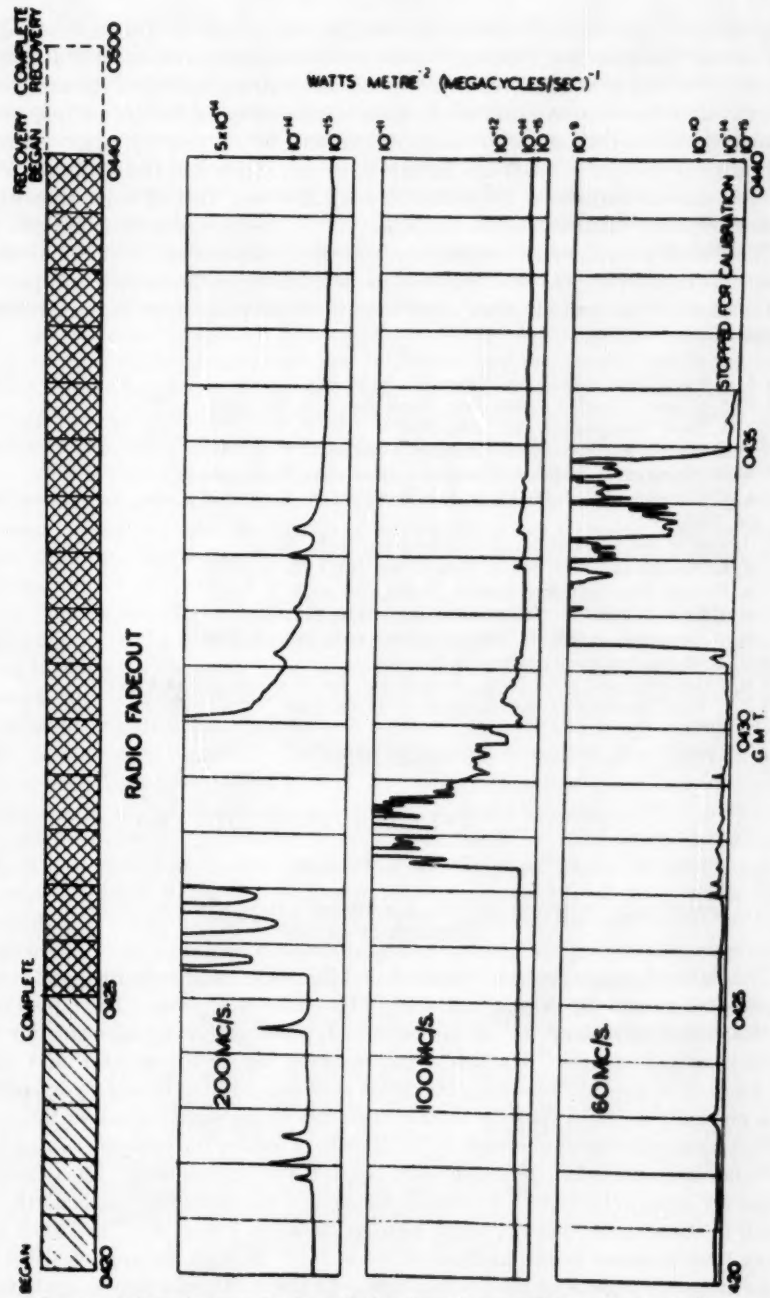


FIG. 4.—A large increase in the radiation from the sun observed in Australia (R. Payne-Scott, D. E. Yabsley, and J. G. Bolton, *Nature*, **160**, 256, 1947).

happenings. The source of such an increase has been observed to move outward in the corona,[13] and the peak frequency of the radiation has been shown to drift downward.[14] Since at a certain level in the corona the electron density will be such as to prevent all radiation above a certain frequency from escaping, the drift in frequency combined with a model of the corona gives a velocity for the exciting agency moving outward. Velocities of 1,000 km/sec and more are found, and the similarity with possible auroral particles is often mentioned. A second kind of burst gives even higher frequency drifts and hence velocities—up to one-fifth the velocity of light.

There is at present much controversy over the interpretation of these observations. It does seem clear however, that radiation in the radio-wave-length part of the spectrum is particularly sensitive to conditions and phenomena in the earth-sun region.

[1] S. F. Smerd, *Australian J. Sci. Research A*, **3**, 34, 1950.
[2] P. A. O'Brien, *Monthly Notices Roy. Astron. Soc.*, **113**, 597, 1954.
[3] J. W. Firor, *Astrophys. J.*, **123**, 320, 1956.
[4] P. A. O'Brien and E. Tandberg-Hanssen, *Observatory*, **75**, 11, 1955.
[5] W. N. Christiansen and J. A. Warburton, *Observatory*, **75**, 9, 1955.
[6] A. E. Covington, W. J. Medd, G. A. Harvey, and N. W. Broten, *J. Roy. Astron. Soc. Can.*, **49**, 235, 1955.
[7] G. Swarup and R. Parthasarathy, *Observatory*, **75**, 8, 1955.
[8] K. E. Machin and F. G. Smith, *Nature*, **170**, 319, 1952.
[9] A. Hewish, *Proc. Roy. Soc., London, A*, **228**, 238, 1955.
[10] F. Link, *Bull. Cent. Ast. Inst. Czechoslovakia*, **3**, 6, 1952.
[11] R. N. Bracewell and G. W. Preston, *Astrophys. J.*, **123**, 14, 1956.
[12] C. A. Shain, *Australian J. Phys.*, **9**, 61, 1956.
[13] R. Payne-Scott and A. G. Little, *Australian J. Sci. Research, A*, **2**, 214, 1949.
[14] J. P. Wild, *Australian J. Sci. Research, A*, **3**, 399, 1950.

THE SOLAR HYDROMAGNETIC DYNAMO

By E. N. PARKER

ENRICO FERMI INSTITUTE FOR NUCLEAR STUDIES AND DEPARTMENT OF PHYSICS, UNIVERSITY OF CHICAGO

The Babcock magnetograms (Babcock and Babcock, 1955) indicate that the sun possesses a general dipole magnetic field of the order of 1 gauss. The dipole field is observed within about 25° of the poles and, presumably, is responsible for the polar coronal streamers. The field is not observed nearer the equator, and what we know of large-scale solar magnetic fields at lower latitudes is only by inference: it is generally assumed that the bipolar character of the sunspots and of the more diffuse magnetic regions observed by the Babcocks implies the existence of bands of toroidal magnetic field circling the sun beneath the photosphere. Each band is about 30° wide in latitude. Successive bands, and corresponding bands north and south of the equator, have opposite sign, as shown in Figure 4. The bands first make their presence known at about latitude ±40° through the appearance of bipolar magnetic regions at the photosphere. In about 11 years they migrate from latitude ±40° to the equator, where they apparently disappear.

Now the surface of the sun is observed to rotate nonuniformly: the equatorial regions have a higher angular velocity than the polar regions and presumably a higher angular velocity than the solar interior; we would expect that beneath the photosphere there is a large-scale shearing of the solar gas. In addition to the large-scale shearing, there is the solar convective zone extending from the base of the photosphere down to a depth of the order of 100,000 km.; as a consequence of the over-all rotation of the sun, we may reasonably suppose that the rising and falling convective cells in the convective zone will be cyclonic. We see, then, that beneath the solar photosphere we have the two essential conditions for a regenerative hydromagnetic dynamo discussed by Elsasser,[1] viz., cyclonic convection and large-scale shearing in an electrically conducting fluid.

It can be shown[2] that when the axes of the cyclonic convective cells are oriented more or less radially about a common center, as they presumably are in the core of the earth, they result in an essentially stationary poloidal dipole field, as the geomagnetic field. In the sun, however, the cyclonic convective cells occur in a relatively thin spherical shell rather than filling a sphere, and, from the point of view of neighboring cells, the curvature of the shell may be neglected and the cell axes taken as parallel; the problem reduces essentially to the effect of parallel cyclonic convective cells throughout a rectangular space in which there is large-scale shearing. As we shall see, the resulting fields are migratory rather than stationary.

To fix ideas, we suppose that the axes of the cyclonic convective cells are in the x-direction, and the shearing motion in the z-direction. We assume that initially there is present a magnetic band consisting of a field $\boldsymbol{B}_t$ in the z-direction. The cyclonic convective cells coming up through $\boldsymbol{B}_t$ carry the lines of force into an upward bulge, which is then twisted into a plane perpendicular to $\boldsymbol{B}_t$, to result in a loop, as shown in Figure 1. Diffusion ultimately severs the loop from $\boldsymbol{B}_t$ and merges neighboring loops (Fig. 2).

The magnetic field of the loop produced by a single convective cell is most conveniently represented in terms of a vector potential $\boldsymbol{A}^{(0)}$ in the z-direction. Physically it is obvious that the average rate of generation of $\boldsymbol{A}^{(0)}$ at a given point in space is proportional to $\boldsymbol{B}_t$ and to the rate of appearance and size of the cyclonic convective cells, which we represent by Γ. Thus, including the usual term $\nabla^2\boldsymbol{A}^{(0)}/\mu\sigma$ (mks units) to take account of the diffusion of the field by the electrical resistivity $1/\sigma$ of the fluid, we write

$$\frac{\partial \boldsymbol{A}^{(0)}}{\partial t} = -\boldsymbol{e}_z \, \Gamma \, \boldsymbol{B}_t + \frac{1}{\mu\sigma} \nabla^2 \boldsymbol{A}^{(0)}, \tag{1}$$

where $\boldsymbol{e}_z$ is a unit vector in the z-direction. The minus sign is necessary, since we have chosen Γ to be positive for a cyclonic convective cell executing a right-hand screw motion. Equation (1) may be derived in a formal way from the magnetic field produced by individual cyclonic convective cells interacting with $\boldsymbol{B}_t$, taking into account the random appearance of the cells in time and space.

The same cells that interact with $\boldsymbol{B}_t$ to generate $\boldsymbol{A}^{(0)}$ will also interact with the y-component of $\nabla \times \boldsymbol{A}^{(0)}$ to generate a new field $\boldsymbol{A}^{(1)}$, so that

$$\frac{\partial \boldsymbol{A}^{(1)}}{\partial t} = + \boldsymbol{e}_y \, \Gamma \, \frac{\partial \boldsymbol{A}^{(0)}}{\partial x} + \frac{1}{\mu\partial} \nabla^2 \boldsymbol{A}^{(1)} \tag{2}$$

The cells will interact with $\boldsymbol{A}^{(1)}$ to form a field $\boldsymbol{A}^{(2)}$ according to

$$\frac{\partial \boldsymbol{A}^{(2)}}{\partial t} = -\boldsymbol{e}_z \, \Gamma \frac{\partial \boldsymbol{A}^{(1)}}{\partial x} + \frac{1}{u\sigma} \nabla^2 \boldsymbol{A}^{(2)}, \tag{3}$$

etc.

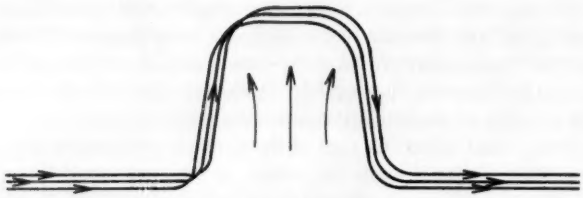


Fig. Ia

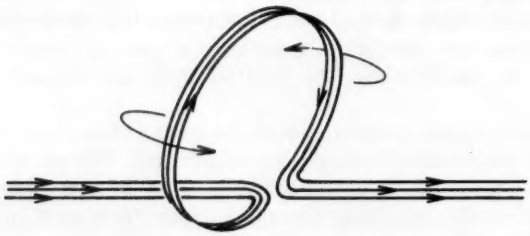


Fig. Ib

FIG. 1.—*a*, Schematic diagram of the upward bulging of the lines of force of the toroidal field by a rising convective cell; *b*, twisting of the bulge to form a magnetic loop with nonvanishing projection on the meridional plane.

Finally, the large-scale shearing $\boldsymbol{e}_z v(x,y)$ of the fluid will shear the x and y-components of the above fields to regenerate $\boldsymbol{B}_t$ according to

$$\frac{\partial \boldsymbol{B}_t}{\partial t} = \nabla \times \left(\boldsymbol{v} \times \sum_{n=0}^{\infty} \nabla \times \boldsymbol{A}^{(2n)} \right) + \frac{1}{\mu\sigma} \nabla^2 \boldsymbol{B}_t$$

$$= + \boldsymbol{e}_z (\nabla v) \times \sum_{n=0}^{\infty} \nabla A^{(2n)} + \frac{1}{\mu\sigma} \nabla^2 \boldsymbol{B}_t \tag{4}$$

Figure 3 is a schematic drawing of the process described by equations (1)–(4). The plus signs represent the lines of force of the band $\boldsymbol{B}_t$ coming out of the plane of the paper. The small circles represent the small loops generated from $\boldsymbol{B}_t$ by the cyclonic convective cells. The large loop represents the result of the diffusive coalescence of the small loops. We suppose that the large-scale shearing results in the fluid near the top of the diagram moving out of the paper relative to the fluid at the bottom. Thus the large loop is drawn out and tipped, so that the right-hand

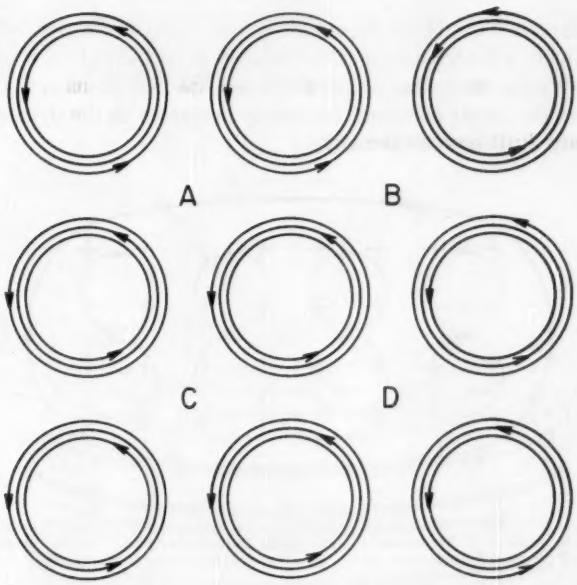


Fig. 2a

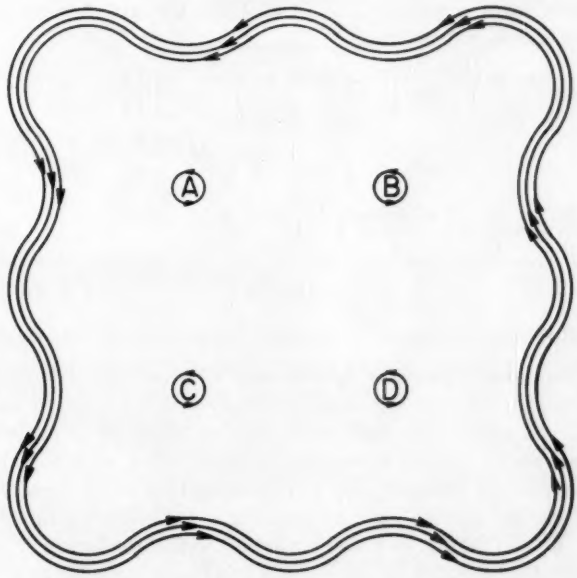


Fig. 2b

FIG. 2.—*a*, Schematic drawing of a number of magnetic loops in a meridional plane, following their formation from the toroidal field by cyclonic convective cells; *b*, schematic large-scale poloidal field formed by interdiffusion and coalescence of loops.

side has a component out of the paper and the left-hand side a component into the paper, resulting in a reinforcement of the band of $\boldsymbol{B}_t$ on the right and a cancellation on the left. We see, then, that we should expect the final position to be somewhat to the right of the initial position; for steady operation of the dynamo we would expect a steady drift toward the right.

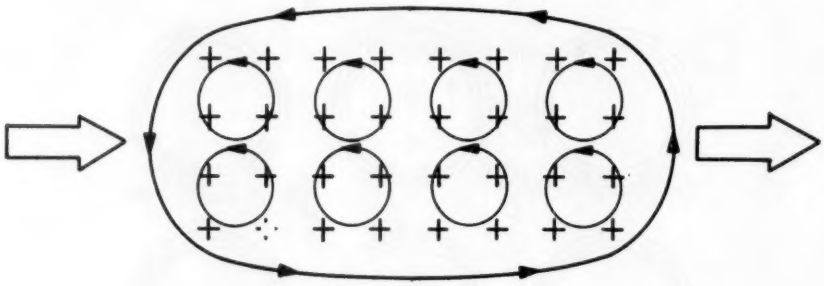


FIG. 3.—Schematic representation of meridional section (xy-plane) through a migratory dynamo: the plus signs represent lines of flux of $\boldsymbol{B}_t$ out of the plane of the paper; the small circles represent the small magnetic loops or rings formed from $\boldsymbol{B}_t$ by cyclonic convective cells; the large loop represents the resultant of the small loops; the large-scale shearing of the fluid is out of the paper at the top and in at the bottom, resulting in the regeneration of $\boldsymbol{B}_t$ in a position displaced from the original in the direction of the open arrows.

Formal solution of equations (1)–(4) confirms the above physical argument. A particularly simple solution is

$$\boldsymbol{B}_t = + \boldsymbol{e}_z B_0 \left(\frac{\partial v/\partial x}{k\Gamma}\right)^{1/2} \exp\left\{\left[\pm\left(\frac{k\Gamma\partial v/\partial x}{2}\right)^{1/2} - \frac{k^2}{\mu\sigma}\right]t\right\} \times \cos\left[\mp\left(\frac{k\Gamma\partial v/\partial x}{2}\right)^{1/2} t + ky\right], \quad (5)$$

$$\boldsymbol{A}^{(0)} = \mp \boldsymbol{e}_z \frac{B_0}{k} \exp\left\{\left[\pm\left(\frac{k\Gamma\partial v/\partial x}{2}\right)^{1/2} - \frac{k^2}{\mu\sigma}\right]t\right\} \times \cos\left[\mp\left(\frac{k\Gamma\partial v/\partial x}{2}\right)^{1/2} t + ky + \frac{\pi}{4}\right], \quad (6)$$

$$\boldsymbol{A}^{(n)} = 0 \qquad (n \geq 1),$$

which we recognize as a traveling *dynamo wave* with velocity $(\Gamma\partial v/\partial x/2k)^{1/2}$.

The amplitude of the wave varies exponentially with time, and, if we assume that the electrical conductivity is sufficiently high to allow us to neglect $k^2/\mu\sigma$, the amplitude changes by a factor of $\exp(2\pi)$, or about 535·4, each period of the wave. If $\Gamma\partial v/\partial x$ is a positive quantity, the wave propagating in the negative y-direction decreases by $\exp(2\pi)$ and the wave propagating in the positive y-direction increases by $\exp(2\pi)$; thus in nature we would not expect to find the former.

Cyclonic rotation vanishes on the equator and has opposite signs in the northern and southern hemispheres; thus in each hemisphere the dynamo waves migrate toward the equator. From equations (5) and (6) we see that the ratio of $|\boldsymbol{B}_t|$ to $|\nabla \times \boldsymbol{A}^{(0)}|$ increases as $\Gamma^{-1/2}$ as the wave approaches the equator; it follows that by

the time the wave has propagated sufficiently far to possess dense fields it will be nearing the equator and $\boldsymbol{B}_t$ will be the dominant field. Figure 4 is a schematic representation of the general solar migratory field.

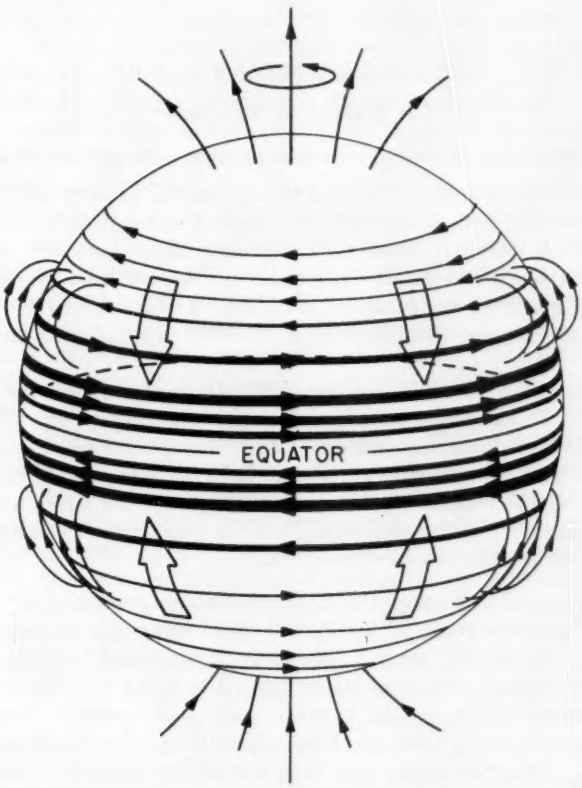


FIG. 4.—Suggested configuration of migratory solar toroidal fields beneath the photosphere. The poloidal loops are indicated near the limb.

As a consequence of the mutual repulsion of magnetic lines of force and the general tendency of a band of magnetic flux to expand, the gas pressure and density within a toroidal band tend to be less than outside. Thus regions of dense field tend to float upward through the fluid, resulting in what has been called *magnetic buoyancy*.[3] When the toroidal fields $\boldsymbol{B}_t$ of the dynamo wave have become sufficiently intense, perhaps several hundred gauss, one might expect that a rope of flux may occasionally manage to float up to the photosphere to form a re-entrant loop. The two intersections of the rope with the photosphere are observed as bipolar magnetic regions. If for some reason the core of the rope should become cool, it has been shown that static equilibrium requires that the diameter of the rope decrease with increasing height in the solar gases, presumably accounting for the very intense magnetic fields found in sunspots.

[1] W. M. Elsasser, in "Geophysics Symposium on Solar and Terrestrial Relationships," these PROCEEDINGS, **43**, 14–24, 1957.
[2] E. N. Parker, *Astrophys. J.*, **122**, 293, 1955.
[3] E. N. Parker, *ibid.*, **121**, 491, 1955.

THE TERRESTRIAL DYNAMO

BY WALTER M. ELSASSER*

DEPARTMENT OF PHYSICS, UNIVERSITY OF UTAH, SALT LAKE CITY, UTAH

It has been known since Gauss that the application of potential theory allows one to separate the geomagnetic field as it appears at the earth's surface into two components. A relatively small part originates above the earth (in the ionosphere and beyond), whereas the largest part is derived from sources inside the earth. Only the latter component of the field will be considered here. One may summarize[1] the detailed characteristics of this field in terms of three outstanding properties:

a) The immense variability of the field both in space, over the earth's surface, and in time, exhibiting periods of the order of a fraction of a century. The secular variation studied in detail by Vestine[2] shows numerous "hills and dales" which move about, grow, and decay, often in but a few decades. Even the strength of the main dipole term is not constant. Somewhat over a century has elapsed since Gauss initiated its first determination; the magnitude of the dipole moment has in that time decreased by over 5 per cent.

b) Analysis of the observed field indicates that the sources of the field interior to the earth's surface are in all likelihood located as deep as the boundary of the liquid core, if not slightly deeper. This may be concluded from the convergence of the spherical-harmonic series representing the field.[1] While one cannot, of course, rigorously determine the location of the field's sources, the assumption that they are essentially below the boundary of the core is in agreement with all known data. In any event, no more than a small fraction of the sources could be assumed to lie above that boundary without contradicting the data. This being established, it becomes in turn desirable to extrapolate the geomagnetic field from the surface of the earth to the boundary of the liquid core on assuming that there are no sources in the intervening mantle. The accuracy of such a projection is severely restricted by the limited precision of the geomagnetic observations; such extrapolation in detail, however, as seems feasible[3] leads one to the conclusion that the topography of the magnetic field at the boundary of the core is rather steep and rugged; one might estimate that the root-mean-square strength of the nondipole field there is perhaps as large as, say, nearly half the root-mean-square strength of the dipole field.

c) The motion, growth, and decay of the features of the geomagnetic secular variation mentioned in paragraph *a* is quite irregular, with one conspicuous exception. There is a mean uniform displacement of the over-all secular pattern from east to west, known as the "westward drift" and amounting in magnitude to 0.18° per year.[4] (This corresponds to a linear velocity of 0.03 cm/sec, which seems to be

typical for the speed of the motions in the earth's core.) The analysis of this phenomenon[4] shows that it represents a differential rotation of the top layers of the liquid core relative to the superimposed mantle. The "westward drift" can be understood[4,5] in terms of a ponderomotive torque of magnetic origin acting between mantle and core, the whole system forming essentially a unipolar induction motor. The phenomenon points unambiguously to the existence of magnetic fields in the core, of the type required by the dynamo theory. Finally, the astronomically observed irregular variations in the rate of rotation of the earth are in all likelihood caused[5] by secular variations of the field near the boundary of the core which make the unipolar induction motor run at a slightly uneven rate.

The interpretation of the extensive observational data mentioned in the foregoing leaves little doubt that the secular variation of the geomagnetic field is due to a mutual coupling, by means of electromagnetic induction, of the magnetic field and the fluid motions in the earth's liquid, metallic core. Conversely, we have here a very strong corroboration of the existence of a liquid core as proposed originally by seismologists on altogether different grounds.

The dynamo theory[6] goes one step further. It assumes that the entire magnetic field is ultimately produced and maintained through fluid motions in the core. Before entering into a description of the dynamo mechanism, we might briefly mention possible alternatives. Disregarding the possibility of ferromagnetic materials inside the earth, which is no longer taken seriously by any workers in the field, we seem to be left with only two possibilities. One of these is the existence of a "fundamental" magnetic field accompanying the rotation of all macroscopic bodies, as sometimes postulated in the past. This may now be safely ruled out on the basis of our more extended knowledge of solar and stellar magnetism (of which more below). The second possibility is represented by electromotive forces setting up electric currents in the core. There can be no doubt that any such current system would be radically modified by the motions prevailing in the core and that the ultimate magnetic field would look altogether different from that generated by emf's in a solid core. Now it may also be shown[6] that eddy motions tend to reduce the effect of the emf-produced currents as compared to a quiescent core; moreover, in bodies larger than the earth's core, such as the sun and stars, the emf-induced currents are altogether negligible. The *dynamical* problem of magnetic fields sustained by fluid motions alone, therefore, ought to have a solution on grounds of the astrophysical evidence. This is the problem to which we address ourselves here, referring, however, to the special physical circumstances that prevail in the earth's core. Any assumptions regarding the action of emf's inside the earth are of necessity highly speculative, and, for the reasons indicated, this mechanism of producing currents, although it cannot be altogether excluded, is likely to be of little importance. If future research bears out the present indications[7] that the polarity of the earth's field has reversed itself occasionally during geological history, a pure dynamo theory would seem to provide the only adequate description of the mechanism generating the observed field.

The generators operating in our power stations show that self-sustaining dynamos are indeed possible. But these machines contain commutators and have withal a very complicated topological structure. The question then arises: Is a dynamo possible inside a singly connected volume of conducting fluid such as the

earth's core? The answer is in the affirmative; we may show that the complicated system of wiring in the technical machine can be replaced by suitable patterns of *motion* inside the fluid. To understand better how this is possible, we must consider the effect which fluid motions in a conductor have upon a magnetic field existing in the same conductor. The behavior of the fluid depends on a dimensionless parameter called the "magnetic Reynolds number,"

$$R_m = \mu\sigma LV,$$

where μ is the magnetic susceptibility, σ the electrical conductivity of the fluid, L a representative length, and V a representative velocity. It may be shown that R_m connects the free decay of electrical currents in the conductor to the periods of mechanical fluid motions, approximately $R_m = \omega_{\text{mech}}/\omega_{\text{el}}$. Here ω_{el} is a measure of the reciprocal free-decay time of the electromagnetic field, and ω_{mech} may be taken as a typical component of the spectrum of the fluid motions. It is known from electromagnetic theory that $\omega_{\text{el}} \sim L^{-2}$: in a large conductor the free-decay times become exceedingly large. Simple calculations show that in the earth's core, and still more so in bodies of cosmic fluids of larger size, we have $R_m \gg 1$. Physically, this means that the fluid can inductively interact with the field before the latter has had time to decay spontaneously. Under these circumstances one can prove, as first shown by Cowling, that

$$\frac{d}{dt}\int B_n dS \Longrightarrow 0$$

where the right-hand side goes to zero as R_m^{-1}. Here d/dt is the substantial derivative in the usual hydrodynamical sense, that is, it refers to the change which occurs as the surface of integration moves along bodily with the fluid. B_n is the component of the magnetic field vector normal to the surface of integration. The content of this formula may be stated by saying that the magnetic lines of force are bodily attached to the fluid particles, and they become deformed correspondingly as the fluid displaces them. A *velocity shear* field is readily seen to lead either to a bundling-together or to a thinning-out of the lines of force. In the former case the total energy of the magnetic field is increased at the expense of the kinetic energy of the fluid, and we have amplification; in the second case the field energy is decreased. It has been suggested by several authors that if an arbitrarily small seminal field exists in a fluid, then the random motions of turbulence will, on statistical grounds, lead to an increase of the field until a random macroscopic field of finite root-mean-square magnitude has been created. This may account for the existence of some magnetic fields in the universe, but it does not constitute a dynamo of the type required, inasmuch as the observed terrestrial, solar, and stellar fields, while exhibiting pronounced random components, are on the whole reasonably regular. It is therefore necessary to consider *ordered* motions if one wants to obtain a satisfactory self-sustaining dynamo. The first idea that occurs to one is to look for patterns of fluid motion that have some degree of symmetry. The simplest such pattern was considered many years ago by Cowling, who assumed that the velocity field is confined to meridional planes. For the sake of consistency, one should assume that the magnetic field is also confined to these planes. An interaction between fluid motion and field exists, but Cowling was able to prove rigorously that such a motion cannot sustain a stationary dynamo.

This result seems capable of considerable generalization;[6] thus motions confined to spherical shells, r = const., also fail to give dynamos, and, more generally, it is not likely that any class of motions in which the trajectories are confined to surfaces, two-dimensional manifolds, can produce a dynamo. These results suggest that actual dynamos have a *low degree of symmetry* of the pattern of the motion rather than a high one. This is corroborated by the fact that the windings and commutators of our technical machines form systems of low symmetry indeed.

It follows from the foregoing that, say, tidal motions in the core could hardly lead to a dynamo mechanism. The motions are most likely convective in origin. This does not necessarily mean thermal convection, but, as Urey has pointed out, small density differences resulting from chemical reactions can equally produce convective motions. The density differences required are so minute that no serious geophysical difficulties are encountered in postulating the existence of convection sustained in the core throughout the age of the earth. One would, however, expect that in the simplest case the trajectories of the fluid particles are confined to meridional planes, and hence by Cowling's argument no dynamo would be possible. To overcome this difficulty, another effect is required which lowers the symmetry of the fluid motions. This is found in the *Coriolis force* produced by the earth's rotation. It is well known that motions in the atmosphere and ocean are completely dominated by the Coriolis force, and, owing to the smaller velocities, this is even more the case for the motions in the earth's core. There is every reason to believe that the naturally occurring dynamos in the earth as well as in stars are characterized by a dominant Coriolis force ("dominant" here means that the Coriolis acceleration in the equations of motion is very large compared to the inertial terms). The hydromagnetic dynamo model here discussed is based on the assumption of a preponderant Coriolis force.

It is clear that the mathematical difficulties encountered in solving a nonlinear dynamical problem coupling the fluid motion with the magnetic field are tremendous. There is one redeeming feature, however, which makes it possible to construct a dynamo model that, in spite of the requirement of low symmetry, is not too complicated. In a dynamo, not only does the motion change the field but, conversely, the ponderomotive forces of the magnetic field modify the fluid motion. The physical nature of these forces is well known: they always act in such a way that they tend to "straighten out," to disentangle the lines of force when the latter get contorted as a result of the fluid motion. Now this tendency clearly cannot lead to a pattern resulting in continued amplification, and hence one may expect that *the basic pattern of the motion which is conducive to dynamo action is determined by the mechanical forces alone (in particular, by the Coriolis force)*. It should therefore be possible to construct dynamos in which the fluid motions are first determined on the basis of purely mechanical considerations involving the Coriolis force; the problem of electromagnetic induction, the dynamo problem, can then be based upon a given velocity field. The latter problem, the *kinematical* dynamo problem, should eventually be capable of a reasonably thorough analysis, since the differential equation for the magnetic field is then linear in the field components.

We may note here that in a stationary dynamo, as Bullard[8] remarked, the free decay of the magnetic field cannot be neglected. Stationarity requires that as

much field energy must be lost per unit time as is generated by the deformation of the field lines. In actual fluids electromagnetic decay must be greatly enhanced by the presence of turbulence. Very little is known about this latter effect, but in a crude approximation we might take care of it by introducing a suitably modified electrical conductivity; this will not be discussed any further here.

Let us first assume that the magnetic field has rotational symmetry about the earth's axis. There are two distinct types of such vector fields. In the first type the lines of force are confined to the meridional planes; this type of field is designated as *poloidal*. In the second type the lines of force are circles about the axis; this type of field is designated as *toroidal*. The most general field of rotational symmetry is a linear combination of these two types; in such a field the field lines have a general spiraling shape. Clearly, these symmetry types have nothing to do with magnetism but refer to the geometry of (divergence-free) vector fields. In this terminology we could restate Cowling's theorem by saying that a poloidal fluid motion acting upon a poloidal magnetic field cannot give rise to a stationary dynamo. (It is, by the way, possible to extend the classification of vector fields into poloidal and toroidal ones to fields that do not have rotational symmetry, on using an analysis in terms of spherical harmonics, but we shall not need this generalization here.)

Consider now a fluid motion which is purely toroidal. This means that the particles describe circles about the earth's axis. In general, the angular velocity may change from circle to circle, and we are dealing with a nonuniform rotation. This motion is of interest because it arises when convection occurs in a rotating mass of fluid. Such convection implies radial mixing; this mixing process has a component normal to the earth's axis. The fluid particles carry angular momentum; those that travel inward have an excess of angular momentum relative to their surroundings, while those that travel outward have a deficit. It follows readily that in a steady-state convective regime there must be a gradient of angular velocity in such a way that the inner parts of the convective layer rotate more rapidly, the outer parts more slowly, than would correspond to uniform rotation.

We now apply this result of elementary mechanics to the deformation of a magnetic field by the fluid motion. Suppose we start with a purely poloidal field such as the earth's dipole field, at time $t = 0$. If now the nonuniform rotation acts upon the field, the lines of force are drawn out along circles of latitude (Fig, 1, *a*, where, for the sake of clarity, the nonuniformity of rotation is assumed concentrated in a thin layer indicated by the dashed circle). This is a process of amplification which can be continued indefinitely if the "primary," poloidal field is kept from decaying. The "secondary," toroidal field thus generated increases linearly with time. Its lines of force are circles of latitude, and the field has opposite sign in the two hemispheres (toroidal quadrupole, Fig. 1, *b*). There is little doubt that this particular process of amplification is the main mechanism whereby energy is pumped from the fluid motion into the magnetic field in the naturally occurring dynamos, the terrestrial one as well as the stellar ones. After this process had been proposed by the author, a suitable model was extensively analyzed by Bullard.[8] There is unfortunately no immediate observational evidence for the existence of a toroidal field inside the earth, since the boundary conditions imposed by the electromagnetic field equations require that this particular type of field vanish outside of con-

ductors; it therefore cannot be detected by measurements at the earth's surface. There is, however, very strong indirect evidence, to be found in the westward drift of the geomagnetic secular field mentioned earlier. This phenomenon, interpreted as a differential rotation at the boundary of the core, can be shown, on analysis, not to be due to any purely mechanical effect; furthermore, the magnetic torque required to produce the differential rotation would necessarily vanish in the absence of a toroidal field. The argument is, in fact, rather more stringent than would appear from this brief description; it leads to the view that the observed westward drift makes the existence of a toroidal field in the earth's core extremely probable if not certain.

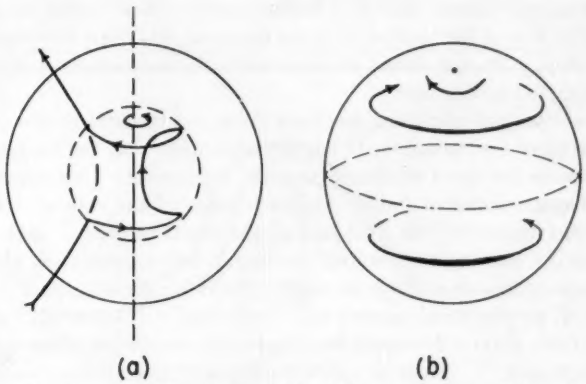


FIG. 1.—*a*, Creation of a toroidal field by nonuniform rotation; *b*, symmetry of the toroidal field.

The nonuniform rotation of the core provides a powerful means of amplification for the toroidal field, but this system does not constitute a dynamo. The reason is that the primary, poloidal field will decay exponentially; and so the induced toroidal field, no matter how large it may become during an intervening period, must eventually decay (needless to say, this decay is exponential with time, while the amplificatory process is linear). Now if we could find a reverse process which would amplify the poloidal field using the toroidal field as the "primary," we would have constructed a dynamo with a two-stage feedback cycle which can maintain itself. The next problem is then to identify this second step, which for simplicity will be designated as the "feedback process." It is easy to show that a fluid motion which would give rise to this process cannot have rotational symmetry about the earth's axis. In fact, assume first that the velocity field is toroidal. Since the velocity and the (toroidal) magnetic field are now everywhere parallel, there is no interaction between them, and the field remains unchanged. Again, let the fluid motion be poloidal. The net result is clearly only a reshuffling of the circular lines of forces, so that the magnetic field remains toroidal. Proceeding in this way, one proves rigorously that no rotationally symmetrical fluid motion can produce a poloidal field from a toroidal primary.

The problem of nonsymmetrical fluid motions has been tackled in two ways. The first approach was developed independently by Bullard and Gellman[9] and by

Takeuchi and Shimazu.[10] In order to find a suitable motion, these authors sort out the vector fields according to the vectorial normal modes of the sphere. They find that the simplest mode which is likely to give feedback is one which is derived from the Legendre function P_2^2. This flow pattern may be geometrically characterized as consisting essentially of four "streams" in the radial direction. The axes of these streams lie in the equational plane; in two of the streams the motion is outward, in the other two, at right angles to the former, the motion is inward. This flow pattern does not in itself provide feedback, but a linear combination of it with the nonuniform rotation already described seems to have the properties required for feedback. The calculations needed to obtain the induced magnetic fields are very laborious and involve series in terms of the spherical-harmonic normal modes. Unfortunately, these series oscillate rather strongly in function of the number of terms involved, and their convergence has not been established. Thus it is not yet certain whether a satisfactory dynamo model can be constructed in this way.

An alternate way of obtaining feedback from the toroidal to the poloidal field has been described by Parker.[11] It has the advantage that the model of fluid motion is chosen on the basis of simple *physical* arguments. In a convective mass, consider a rising column of fluid. At the bottom of the column there must be horizontal convergence of the fluid and at the top divergence. But the Coriolis force deflects the converging flow and converts it into a circulation about the axis of the stream. This circulation is counterclockwise on a map of the northern hemisphere or, as the geophysicists say, "cyclonic." (Conversely, the action of the Coriolis force upon a divergent flow leads to a circulation of the opposite sense, called "anticyclonic.") Thus in a rising eddy we have cyclonic circulation combined with a vertical, outward displacement, the net result being a helical motion of the fluid particles. Consider now what happens if this type of motion deforms the toroidal magnetic field. In Figure 2 a strand of the toroidal field is shown which is lifted and at the same time twisted cyclonically. The net result is clearly the generation of a magnetic loop which has (for a suitable angle of twist) a nonvanishing projection upon the plane normal to the direction of the toroidal field that is, upon a meridional plane.

Assume now that there are many such convective eddies. Each eddy produces what we may call a local magnetic feedback loop; the field lines of any such loop have a nonvanishing poloidal component, that is, nonvanishing projections upon the meridional planes. As time goes on, the effect of free decay (expressing itself in a diffusion of field lines through the fluid) will make these field lines coalesce into an over-all poloidal field. Parker[11] has furthermore shown that it is easy to obtain in this way a poloidal field in which the dipole component predominates, as is observed in the earth. It is only necessary to assume that loop formation goes on at some depth but does not extend entirely to the surface of the core. The layer near the surface which is devoid of feedback loops tends to suppress the higher harmonics of the poloidal field on the outside of the core.

Some comments on this feedback mechanism are in order. We note that this system does not constitute a stationary dynamo. The convective eddies, akin to turbulence, will be quite irregular, and there seems to be little resemblance in this to a stationary flow pattern. We have seen before that the feedback mechanism

must involve deviations from rotational symmetry; thus any effective flow pattern is of necessity complex. On introducing the notion of a set of *local* eddies that provide feedback in a statistical way, we are led to a dynamo model which is stationary *in the mean*. On the other hand, on averaging over these eddies, including a time average, we obtain a model which in the mean has rotational symmetry. The deviations from rotational symmetry have been relegated to the statistical correlations in local eddies. The latter will now be discussed somewhat more closely.

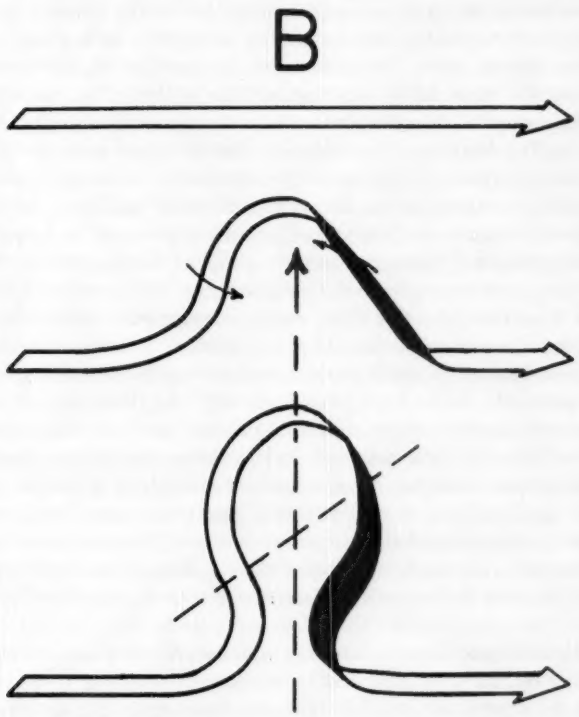


FIG. 2.—Creation of a magnetic feedback loop by combination of rising and cyclonic motion.

Start for simplicity with a slab of fluid in which convection takes place normal to the boundaries and the axis of rotation is also normal to the boundaries. There must, of course, be as much descending as ascending matter. Now, if a vertical stream is formed by convergence, there should always be a cyclonic circulation. Thus in the simplest model a cyclonic twist would be accompanied by upward motion in half the streams and by downward motion in the other half. Consequently, in this simplified picture there would be magnetic loops of opposite signs in equal numbers. Clearly, then, the actual motion of the fluid must be more complicated if the feedback process is to operate. We must postulate that there is a

preferential correlation of upgoing motion with cyclonic circulation and of downgoing motion with anticyclonic circulation. If the model adopted for the individual eddies is not too complicated, this may be shown to be equivalent to a positive correlation between the vertical component of the velocity and the vertical component of the vorticity. (In case of a correlation of opposite sign the feedback would be degenerative rather than regenerative.) In a spherical model the correlation would of course apply to the corresponding radial components. Now what is the nature of this correlation? This cannot be said conclusively at present, but some interesting arguments can be propounded.

Experience shows that such a correlation is universally present in the motions of the atmosphere: cyclones are invariably associated with rising air and anticyclones with sinking air. The dynamical background of this behavior is not very clear, but the most likely assumption here is that of a connection with the vertical asymmetry of the atmospheric convective regime. The cyclones owe their origin to the heating of a relatively shallow layer near the ground which then converges and rises in a few concentrated regions. The corresponding cooling in the upper air is caused by radiation and is rather uniform; the sinking anticyclonic motions extend over large areas, and they are very far from being mirror images of the cyclones. Quite possibly the cause of the required correlation must be sought, then, in asymmetries of the underlying heat-transfer processes. The dynamics of Coriolis-controlled fluid motions being still extremely obscure, we can do no more than intimate the direction in which a solution might be found. Next, one would have to apply similar arguments to convection in the earth's core. It is plausible, as we have tried to show,[6] that there is also an asymmetry in the regime of transfer, the cyclones originating probably from the rising of a relatively thin layer of light material that forms on top of the presumably solid "inner core" (whose diameter, from seismic observations, is about one-third that of the whole core). Next, if one wishes to apply the same arguments to convective layers in the sun and stars, the asymmetry of thermal convection might be related to the extremely rapid decrease of density radially outward (in the sun, for instance, the density decreases by a factor of 2 over a distance of a few hundred kilometers).

Clearly, these arguments are far from conclusive; they are merely suggestive. They should eventually be justified by a more painstaking dynamical analysis. There is, on the other hand, the fact that prolonged study has not brought to light any other physically founded mechanism that could readily and simply produce feedback from the toroidal to the poloidal field. It is therefore desirable to *postulate* that the described correlation between vertical velocity and circulation, or else vorticity, exists in the earth's core. Furthermore, since stellar convective layers are certainly more similar to the earth's atmosphere than the fluid core is, one may postulate that the same correlation prevails in the sun and stars. Thus we have a tentative scheme of a *general* hydromagnetic dynamo mechanism which is in essence a two-stage feedback process: A toroidal field is produced from an existing poloidal field by the nonuniformity of rotation, this being the process in which most of the magnetic field energy is created out of kinetic energy of the fluid. Given this toroidal field, the poloidal field is then regenerated through feedback loops that result from the postulated correlation of velocity and vorticity com-

ponents. Thus the gyroscopic forces enter essentially into *both* stages of the feedback cycle. We are still far from a complete mathematical analysis of this dynamo model, although such analysis might seem within the realm of the possible.[6]

A rather unexpected corroboration of this view on the controlling influence of the Coriolis force has recently been obtained from astrophysics. Babcock[12] has now identified some thirty-odd magnetic stars. The conditions for the detection of a magnetic star are rather stringent: the object must be fairly close to the earth and bright, the spectral lines must be rather sharp, not too much broadened by rotation, and, finally, the linear average of the line-of-sight component of the field over the star's disk must not be much less than 1,000 gauss, since the precision of the measurements is only about ± 500 gauss. The magnetic stars are concentrated in the middle of the spectral sequence; the majority are A stars, and some others are in subsequent classes. From a statistical analysis of his data, Babcock concludes that, given the difficulties of observing the stellar magnetic fields at all, a large fraction of all the stars in these classes are likely to be magnetic. Now according to our dynamical notions, a hydromagnetic dynamo requires the presence of both convection and rotation. It is known that the earliest-type stars do not have convective layers; a convective outer shell develops only in the A stars and becomes progressively deeper in the later types. On the other hand, the early types as a rule rotate rapidly, and the later types do, on the average, rotate progressively more slowly. Thus, if the dynamo mechanism is Coriolis-controlled, one would expect that the incidence of magnetic stars has a maximum at some intermediate point in the stellar sequence where both rotation and convection are appreciable; this is exactly what is observed.

The sun is a relatively late star; it rotates relatively slowly but has an outer convective layer whose depth is of the order of about one-fifth of the solar radius. There seems to be very little doubt that sunspots owe their origin to strands of the solar toroidal field in the convection zone which have broken through the solar surface, the magnetic field having, by some as yet unidentified cause, been strongly concentrated to give the observed values of a few thousand gauss inside sunspots. (The much more extensive magnetic regions observed by Babcock have fields of only a few gauss.) The existence of solar poloidal fields has been inferred from the typical streakiness of the solar corona, especially in the polar regions (though these fields in the outer regions are rather weak, of the order of only 1–2 gauss). There can be little doubt that both poloidal and toroidal fields exist in the sun, and the average symmetry of the field about the solar axis is apparent, indicating again a close connection with the solar rotation. Parker[11] has made a promising beginning in an attempt to explain magnetic fields in the solar convection zone and their variation in the sunspot cycle by means of a two-stage feedback scheme whose basic features are those described above. In one respect the solar poloidal field differs characteristically from the terrestrial one: in the earth the dynamo mechanism seems to fill most of the interior volume of the core, and this leads to a coalescence of the loops into a field in which the dipole preponderates. In the sun, on the other hand, the convective region forms a layer right below the surface, fairly thin compared to the solar radius. One can then anticipate a more complicated structure of the solar poloidal field, and observations seem to bear out that there is no over-all dipole; only in the regions of the polar caps does the dipole term predominate.

There is no question as to the fundamental importance of hydromagnetic fields in the so-called solar activity which determines the multitudinous and complex solar-terrestrial relationships. The study of the geomagnetic dynamo might lead us to a better understanding of the related phenomena on the larger stage of the sun.

* Now at University of California, Scripps Institute of Oceanography, La Jolla, Calif.

[1] W. M. Elsasser, *Revs. Mod. Phys.*, **22**, 1, 1950.

[2] E. H. Vestine *et al.*, *Carnegie Inst. Washington Publs.*, Nos. 578, 580, 1947.

[3] K. L. McDonald, *J. Geophys. Research*, **60**, 377, 1955.

[4] E. Bullard, C. Freedman, H. Gellman, and J. Nixon, *Phil. Trans. Roy. Soc. London, A*, **243**, 67, 1950.

[5] W. Elsasser and H. Takeuchi, *Trans. Am. Geophys. Union*, **36**, 584, 1955; *J. Phys. Earth*, **2**, 39, 1954.

[6] W. M. Elsasser, *Revs. Mod. Phys.*, **28**, 135, 1956. This paper contains an extensive account of the dynamo theory, especially in the more mathematical aspects omitted here. For a review of cosmic magnetism from a more general viewpoint, see W. M. Elsasser, *Am. J. Phys.*, **23**, 590, 1955; **24**, 84, 1956.

[7] S. K. Runcorn, *Advances in Physics* (*Phil. Mag.* Suppl.) **4**, 244, 1955.

[8] E. C. Bullard, *Proc. Roy. Soc. London, A*, **197**, 433; **199**, 413, 1949.

[9] E. Bullard and H. Gellman, *Phil. Trans. Roy. Soc., London, A*, **247**, 213, 1954.

[10] H. Takeuchi and Y. Shimazu, *J. Phys. Earth*, **1**, 1, 57, 1952; *J. Geophys. Research*, **58**, 47 1953.

[11] E. N. Parker, *Astrophys. J.*, **122**, 293, 1955.

[12] H. Babcock and T. Cowling, *Monthly Notices Roy. Astron. Soc.*, **113**, 357, 1953.

ON COSMIC MAGNETIC FIELDS

BY S. CHANDRASEKHAR

UNIVERSITY OF CHICAGO

1. When magnetic fields occur on the cosmic scale—as in interplanetary regions, in stellar atmospheres and in stellar envelopes, in stars and in interstellar space—they have patterns and properties which cannot be deduced by simply extrapolating from terrestrial experience: for, in these instances, the scale is the essence of the phenomenon, and the scale transcends terrestrial experience.

2. The principal characteristic of magnetic fields pervading large linear dimensions is their relative permanence. This is so because the mean life of a magnetic field with a scale length L, apart from a numerical factor, is given by σL^2, where σ is a measure of the electrical conductivity of the medium; and for values of L appropriate in the astronomical contexts this can be very long indeed. An immediate consequence of the long mean lives is that media of cosmical dimensions have attributes one normally associates with infinite electrical conductivity. Therefore, we may picture the lines of force, even as Kelvin and Helmholtz pictured the vortex lines in the ether, as indestructible and as effectively "frozen" in the medium.

3. Under the circumstances described, the magnetic field can change only by the lines of force being dragged about by the fluid. Conversely, the magnetic field also reacts on the motion through the operation of the Lorentz force

$$\boldsymbol{L} = j \times \boldsymbol{H} = \frac{\mu}{4\pi} \operatorname{curl} \boldsymbol{H} \times \tag{1}$$

(where j denotes the current and μ the coefficient of magnetic permeability). The equations of motion, in fact, are

$$\frac{\partial \boldsymbol{H}}{\partial t} = \operatorname{curl} (\boldsymbol{v} \times \boldsymbol{H}), \tag{2}$$

$$\rho \frac{\partial \boldsymbol{v}}{\partial t} + \rho(\boldsymbol{v} \cdot \operatorname{grad}) \boldsymbol{v} = -\rho \operatorname{grad} p + \operatorname{grad} V + \rho \boldsymbol{L}, \tag{3}$$

$$\frac{\partial \rho}{\partial t} + \operatorname{div} (\rho \boldsymbol{v}) = 0 \qquad \text{and} \qquad \operatorname{div} \boldsymbol{H} = 0. \tag{4}$$

In equation (3) V denotes the gravitational potential, p the pressure, and ρ the density.

4. In virtue of equations (2) and (3), the norm of behavior when the energy density of the magnetic field is comparable to, or exceeds, the energy density of the mass motions is quite different from what it would be if this were not the case. Thus, if the magnetic field dominates and determines the motions, we may expect the pattern of the field and of the motions to evolve along one of two directions: *Either* the magnetic field gets altered in such a way that its influence on the motions is progressively reduced and a stationary static state is eventually reached, *or* the motions are amplified at the expense of the field until some kind of a balance in the interaction between the two fields is reached.

5. For the attainment of the first of the two end states envisaged, it is necessary that, in the limit,

$$\boldsymbol{L} = \frac{\mu}{4\pi} \operatorname{curl} \boldsymbol{H} \times \boldsymbol{H} = 0, \tag{5}$$

or

$$\operatorname{curl} \boldsymbol{H} = \alpha \boldsymbol{H}, \tag{6}$$

where α is a function of position. Magnetic fields having this property are called *force-free*. Their importance in providing possible basic patterns of cosmic magnetic fields was first recognized by Lüst and Schlüter.[1]

6. One obvious way in which condition (5) can be fulfilled is by requiring that

$$j = \frac{\mu}{4\pi} \operatorname{curl} \boldsymbol{H} = 0. \tag{7}$$

This condition that space be current-free and $\boldsymbol{H}$ be derivable from a scalar potential has generally been used to characterize magnetic fields outside material bodies; it is used, for example, in specifying the geomagnetic field outside the earth's surface. However, in most astronomical connections the condition $j = 0$ cannot apply, since, on account of the relatively high electrical conductivity of the stellar material, currents can freely flow in regions of low density: as low as in the corona or even as low as in interstellar space. Therefore, as Lüst and Schlüter have

emphasized, one should seek solutions of equation (6) more general than equation (7).

The case when α is a constant and the magnetic field has symmetry about an axis is of particular interest. The general solution characterizing such force-free magnetic fields has recently been given.[2]

7. A question of considerable practical importance in this connection is the stability of the force-free magnetic fields. This remains to be investigated.

8. A further related question concerns the nature of magnetic fields which can prevail under conditions of hydrostatic equilibrium and which are not force-free. This question resolves itself into finding the conditions for the integrability of the equation

$$\text{grad } p = \rho \text{ grad } V + \boldsymbol{L}. \tag{8}$$

If we should be dealing with media of uniform density, the integrability of equation (8) requires that

$$\text{curl } \boldsymbol{L} = 0. \tag{9}$$

In other words, $\boldsymbol{L}$ should be derivable as the gradient of a scalar function. Prendergast and the writer[3] have recently determined what this condition means for axisymmetric magnetic fields; and Prendergast[4] has constructed a model which satisfies this condition and represents a true equilibrium configuration.

9. As an alternative to the magnetic field becoming force-free, it was also envisaged (cf. sec. 4) that the interaction between the magnetic field and the fluid motion might reach a state of balance. It is customary[5] in these connections to suppose that the state of balance is one of turbulence in which there is an equipartition of energy between the two forms; thus

$$\frac{1}{2}\rho\langle|\boldsymbol{v}|^2\rangle_{\text{av}} = \frac{1}{8\pi}\langle|\boldsymbol{H}|^2\rangle_{\text{av}}. \tag{10}$$

10. For an incompressible medium a state of balance between the magnetic field and the fluid motion need not necessarily be one of turbulence. For in this case the velocity is also a solenoidal vector, and the equations of hydromagnetics (eqs. [2], [3], and [4]) simplify, and in tensor notation take the forms

$$\frac{\partial v_i}{\partial t} + \frac{\partial}{\partial x_j}\left(v_i v_j - \frac{\mu}{4\pi\rho}H_i H_j\right) = -\frac{\partial}{\partial x_i}\left(\frac{p}{\rho} + V + \frac{\mu|\boldsymbol{H}|^2}{8\pi\rho}\right) \tag{11}$$

and

$$\frac{\partial H_i}{\partial t} + \frac{\partial}{\partial x_j}(H_i v_j - v_i H_j) = 0. \tag{12}$$

From these equations it is evident that

$$v_i = \pm H_i\left(\frac{\mu}{4\pi\rho}\right)^{1/2} \tag{13}$$

and

$$p + \rho V + \frac{\mu|\boldsymbol{H}|^2}{8\pi} = \text{constant} \tag{14}$$

represent an exact stationary solution. Moreover, it has been shown that this solution is *stable* for small oscillations.[6]

11. The principal significance of solution (13), apart from its stability, is this: aside from their solenoidal character, we do not need to restrict $\boldsymbol{H}$ or $\boldsymbol{v}$ in any way. This freedom in the choice of $\boldsymbol{H}$ makes one wonder whether magnetic fields, when they occur on the cosmic scale, are not always accompanied by correlated motions.

12. In connection with the remarks in section 11, it is of interest to inquire into the conditions under which the general equations of hydromagnetics (2)–(4) will also allow equation (13) as a stationary solution.

It is clear that equation (2) is compatible with this solution and a stationary state.

Next, by combining equation (3) and the equation of continuity, we have

$$\frac{\partial}{\partial t}(\rho v_i) + \frac{\partial}{\partial x_j}\left(\rho v_i v_j - \frac{\mu}{4\pi} H_i H_j\right) = -\text{grad}\left(p + \frac{\mu |\boldsymbol{H}|^2}{8\pi}\right) + \rho \text{ grad } V; \quad (15)$$

and this equation is also compatible with solution (13) and a stationary state, provided we satisfy the hydrostatic relation

$$\text{grad}\left(p + \frac{\mu |\boldsymbol{H}|^2}{8\pi}\right) = \rho \text{ grad } V. \quad (16)$$

However, the fact that in the stationary state $\rho\boldsymbol{v}$ (in addition to $\boldsymbol{H}$) must also be solenoidal requires that

$$\text{div}\,(\boldsymbol{v}\sqrt{\rho}) = 0 \quad \text{and} \quad \text{div}\,(\boldsymbol{v}\rho) = 0. \quad (17)$$

Both these equations can be simultaneously satisfied only if ρ is constant along the trajectories and $\boldsymbol{v}$ is solenoidal.

13. It is clear that we are still very far from an adequate characterization of cosmic magnetic fields.

[1] R. Lüst and A. Schlüter, "Kraftfreie Magnetfelder," *Z. Astrophys.*, **34**, 263–282, 1954.

[2] S. Chandrasekhar, "On Force-free Magnetic Fields," these PROCEEDINGS, **42**, 1–5, 1956.

[3] S. Chandrasekhar and Kevin H. Prendergast, "The Equilibrium of Magnetic Stars," these PROCEEDINGS, **42**, 5–9, 1956. For a generalization of the theorem proved in this paper, which includes at the same time the theorems on force-free magnetic fields and the so-called law of isorotation, see S. Chandrasekhar, "Axisymmetric Magnetic Fields and Fluid Motions," *Astrophys. J.*, **124**, 232–243, 1956; the theorem in question is included in equation (55).

[4] Kevin H. Prendergast, "The Equilibrium of a Self-gravitating Incompressible Fluid Sphere with a Magnetic Field. I," *Astrophys. J.*, **123**, 498–507, 1956.

[5] E. Fermi, "On the Origin of the Cosmic Radiation," *Phys. Rev.*, **75**, 1169–1174, 1949; A. Schlüter and L. Biermann, "Interstellare Magnetfelder," *Z. Naturforsch.*, **5a**, 237–251, 1950; for a more quantitative discussion see S. Chandrasekhar, "Hydromagnetic Turbulence. I. A Deductive Theory," *Proc. Roy. Soc. London, A*, **233**, 322–330, 1955; S. Chandrasekhar, "Hydromagnetic Turbulence. II. An Elementary Theory," *ibid.*, pp. 330–350.

[6] S. Chandrasekhar, "On the Stability of the Simplest Solution of the Equations of Hydromagnetics," these PROCEEDINGS, **42**, 273–276, 1956.

SOLAR INFLUENCES ON COSMIC RAYS

By Scott E. Forbush

CARNEGIE INSTITUTION OF WASHINGTON, DEPARTMENT OF TERRESTRIAL MAGNETISM, WASHINGTON D.C.

Introduction.—Probably all the established variations with time of cosmic-ray intensity are directly or indirectly due to solar influences. Ionization chambers, because they are relatively simple and reliable, have proved valuable for continuous registration of cosmic-ray intensity over long periods of time. The longest series of continuous observations with several identical instruments of this type is that which has been obtained with the Carnegie Institution of Washington model C Compton-Bennett meters.[1] With these meters, and with the unselfish co-operation of several organizations, continuous data have been obtained for nearly two decades at each of the following places:[2] Godhavn (Greenland), Cheltenham (Maryland), Huancayo (Peru), and Christchurch (New Zealand). Data for shorter periods have been obtained from Teoloyucan and Ciudad Universitaria, Mexico, D.F., and from Climax, Colorado. These ionization chambers are shielded by the equivalent of 12 cm. Pb to screen out local gamma rays, the intensity of which may vary with time.

Origin of μ-Meson Component Measured in Ionization Chambers and Its Seasonal Variation.—In ionization chambers the ionization is normally produced mainly by μ-mesons, which have a mass of about 210 electron masses, unit electronic charge (±), and a lifetime (at rest) of about 2×10^{-6} seconds. These very penetrating μ-mesons arise from the decay of short-lived charged π-mesons (lifetime at rest about 10^{-8} seconds), which in turn result from the interaction of high-energy protons of the primary cosmic-ray beam with atmospheric nuclei. There are other types of reaction with different decay products which need not concern us here. The maximum μ-meson intensity occurs in the region where the pressure is roughly 100 mb., which normally is near an altitude of 16 km. or so. An increase in the height of this pressure level lengthens the path for μ-mesons, so that more of them decay into not very penetrating electrons (and nutrinos) before reaching the instruments at the ground. Thus the seasonal variation in height of the 100-mb. level results in a seasonal variation in cosmic-ray intensity as measured by ionization chambers. The passage of meteorological "fronts" can also alter the height of the 100-mb. level and consequently the apparent cosmic-ray intensity as recorded in ionization chambers (or with Geiger counters). The seasonal waves and meteorological effects proved less interesting than troublesome, for they obscured for a while the more interesting remaining changes in cosmic-ray intensity which are worldwide[3] and result directly or indirectly from solar influences. Since ionization chambers are mainly sensitive to the μ-meson component, it is essential to remark that normally most of the ionization in such detectors arises from primaries of relatively high energy. For example, near sea level, the intensity in ionization chambers near the geomagnetic pole is ordinarily only about 10 per cent greater than at the equator. This means that about 90 per cent of the ionization results from primaries which have enough momentum to arrive at the earth's geomagnetic equator (about 15 Bev/c for protons). The primaries of lower momentum which are not prevented by the earth's magnetic field from reaching the earth, say at or

north of geomagnetic latitude 50°, are ineffective in producing penetrating μ-mesons.

Origin of the Nucleonic Component and Solar-Flare Effects.—The primary cosmic-ray particles, protons or heavier nuclei, also generate a cascade of nucleons (protons and neutrons) in the atmosphere; even relatively low-energy protons, with momenta only great enough (a few Bev/c) to be allowed by the earth's field to reach the atmosphere at geomagnetic latitude 50°, are very effective generators of the nucleonic component. Most of the rare large increases in cosmic-ray intensity which have been accompanied by solar flares or chromospheric eruptions are due to relatively low-energy charged particles coming from the sun.[4] The momenta of most of these particles are too low, as shown by their large latitude effect, to result in production of μ-mesons (through the $\pi \rightarrow \mu$ decay). The large augmentation of ionization observed in ionization chambers during these solar-flare effects results from the nucleonic component, as was shown[4] by the fact that the magnitude of the augmentation increased with altitude at the same rate as that known for the nucleonic component. During the solar flare of November 19, 1949, the total ionization in a Compton-Bennett meter at Climax, Colorado, increased about 200 per cent. At Climax, under 12 cm. Pb, probably no more than 10 per cent of the total ionization is normally due to local radiation originating from the nucleonic component. Assuming that this normal radiation was produced by particles in the same energy band as that responsible for the increase of 200 per cent (at Climax) in ionization which accompanied the solar flare of November 19, 1949, then it was predicted[4] that the rate of arrival of primary particles in that energy band must have increased to at least 20 times the normal value. During the solar flare[5] of February 23, 1956, increases of 20-fold and greater have been reported from detectors sensitive only to the nucleonic component. Thus neutron detectors have many advantages over the ionization chamber for measuring effects which arise from variations in the intensity of the low-energy part of the cosmic-ray spectrum.

Magnetic-Storm Effects.—Figure 1 shows a decrease of daily means of cosmic-ray intensity similar at three stations during the magnetic storm of April 25–30, 1937. The bottom curve indicates the decrease and subsequent gradual recovery of daily means of horizontal magnetic intensity, H, at Huancayo. It will be noted that the rate of recovery of H toward normal is similar to that for cosmic-ray intensity. Figure 2 shows three examples of decreases in daily means of cosmic-ray intensity associated with decreases in horizontal magnetic intensity at Huancayo during the period January 11–31, 1938, which contains what may for convenience be called three separate magnetic storms. While the ratio of changes in cosmic-ray intensity to those in horizontal magnetic intensity is relatively constant throughout any one of these storms, it differs among the three storms. Not all storms are accompanied by decreases in cosmic-ray intensity. The decrease in horizontal magnetic intensity at Huancayo for the magnetic storm beginning August 21, 1937, was accompanied by no detectable decrease in cosmic-ray intensity in the Compton-Bennett ionization chambers. Figure 3 depicts the daily means of cosmic-ray intensity and of horizontal magnetic intensity, H, at Huancayo for 1946. The decrease of several per cent in cosmic-ray intensity between February 3 and 6 occurred before the major depression in H which took place on February 7. There was, however, on February 3 a sudden magnetic commencement, although this was not im-

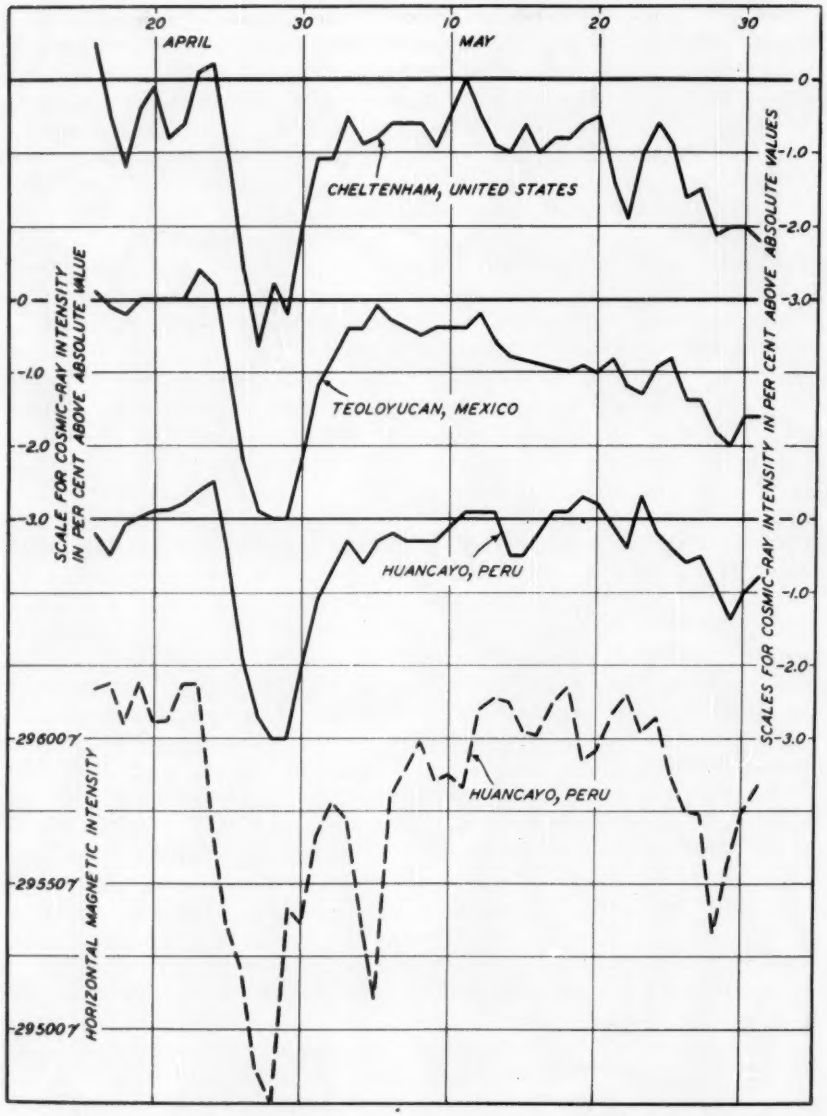


FIG. 1.—Daily means of cosmic-ray intensity and horizontal magnetic intensity, showing effect of magnetic storm of April 25–30, 1937, on cosmic-ray intensity.

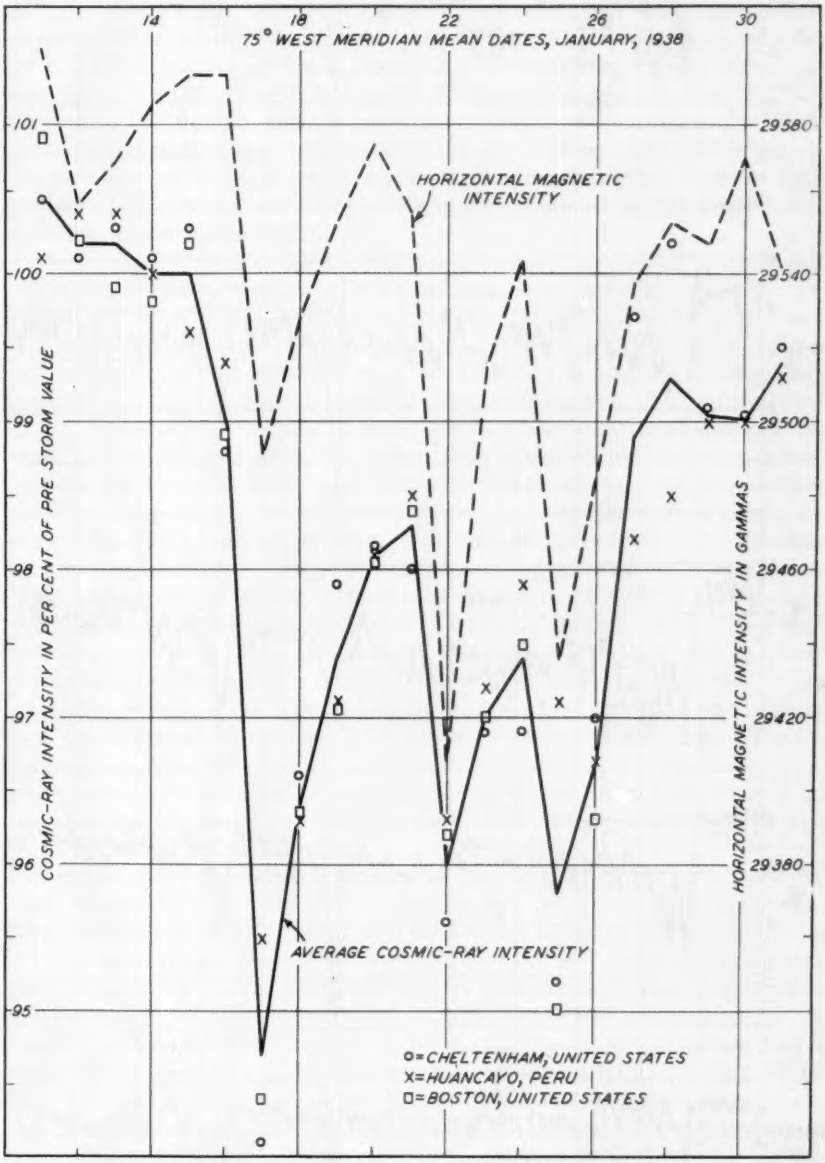


FIG. 2.—Magnetic-storm effects on daily mean cosmic-ray intensity at Boston, United States, Cheltenham, United States, and Huancayo, Peru, and on daily mean magnetic horizontal intensity at Huancayo, Peru.

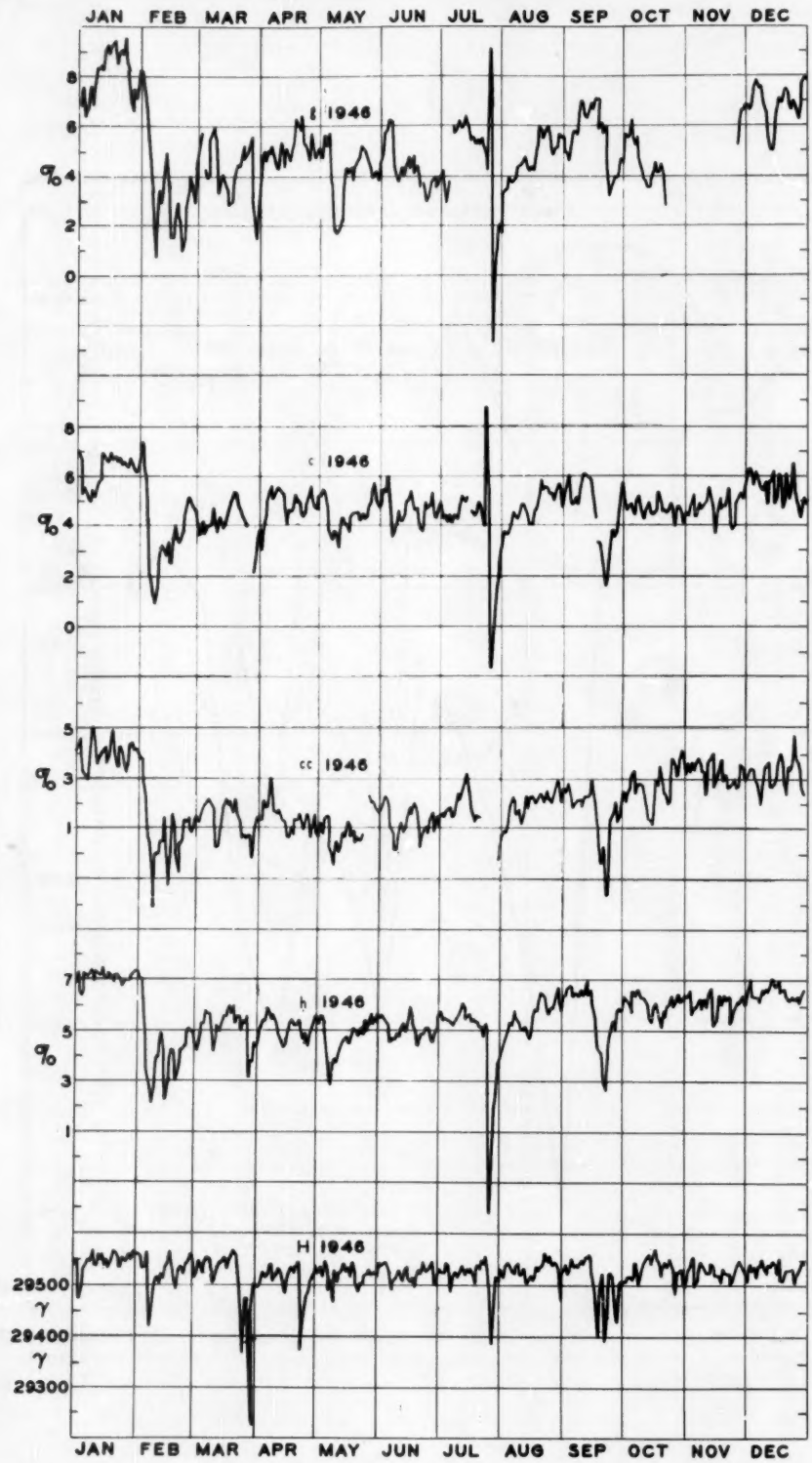


FIG. 3.—Daily means, cosmic-ray intensity, 1946. *g* = Godhavn, *c* = Cheltenham, *cc* = Christchurch, *h* = Huancayo; *H* = horizontal magnetic intensity at Huancayo.

mediately followed by a magnetic storm. This particular decrease in cosmic-ray intensity, which began February 3, 1946, is unusual in that the cosmic-ray intensity remained low for much of the remainder of the year. Incidentally, the large values for July 25 at Godhavn (*g*) and at Cheltenham (*c*) result from the large increase associated with a solar flare on that date. No flare effect was registered at Huancayo, since the charged particles from the sun had insufficient momenta for the earth's magnetic field to allow their arrival at the equator. At Christchurch (*cc*) the flare occurred during a period of a few days when the equipment was not operating. In Figure 3 it is evident that the major changes in daily means of cosmic-ray intensity are world-wide.[3]

Variability of Daily Means.—Figure 3 shows the variations in daily mean cosmic-ray intensity for several stations for a year near maximum sunspot activity. The curves for 1944 in Figure 4 indicate the variations of daily means for 1944 at sunspot minimum. The curve *h 1944* for daily means of cosmic-ray intensity at Huancayo exhibits much less variability than for 1946, and the same is true for the daily means of magnetic horizontal intensity (*H 1944*) at Huancayo. Each of the curves (*g 1944*, *c 1944*, and *cc 1944*) in Figure 8 exhibits greater variability, especially during winter, than does *h 1944*. Moreover, most of this variability is uncorrelated between different pairs of stations. Except at Huancayo, this variability is doubtless due to μ-meson decay effects resulting from changes in the vertical distribution of air mass with the passage of fronts. At Huancayo the variability of daily means in some months of 1944 is not greatly in excess of the inherent "noise level" of the instrument. Thus Huancayo is essentially free of the disturbing effects arising from the passage of meteorological "fronts" which do not occur there. The standard deviation of daily means from monthly means (pooled for each year) of cosmic-ray intensity at Huancayo is shown in Figure 5 for the period 1937–1955. This variability is least in the years of sunspot minima, 1944 and 1954, and increases near years of sunspot maxima. The variability is somewhat less when the five magnetically disturbed days of each month are excluded. This shows that in nearly every year significant cosmic-ray changes occur on at least some of the magnetically disturbed days. In some years much of the variability arises from the 27-day quasi-periodic variation.[2]

Disturbed minus Quiet-Day Difference.—For each month geomagneticians determine the five days when the earth's magnetic field is quiestest and the five days when it is most disturbed. In Figure 6 the average difference in cosmic-ray intensity for the disturbed less that for the quiet days for each month at Huancayo is plotted against the corresponding difference for Cheltenham. These differences are preponderantly negative and correlated between the two stations, showing that the cosmic-ray intensity definitely tends to be less on magnetically disturbed than on magnetically quiet days. The annual means for magnetically disturbed (60 per year) less quiet days (60 per year) are shown in Figure 7 for cosmic-ray intensity at each of three stations (and their average) and for magnetic horizontal intensity. These differences are all negative and vary (algebraically) roughly with the sunspot cycle.

Twenty-seven-Day Variation.—Figure 8 (*A*) shows the variability of the 27-day waves in magnetic activity (American magnetic character figure) and in cosmic-ray intensity at Huancayo (*B*), for a sample of 34 solar rotations. The maxima of the

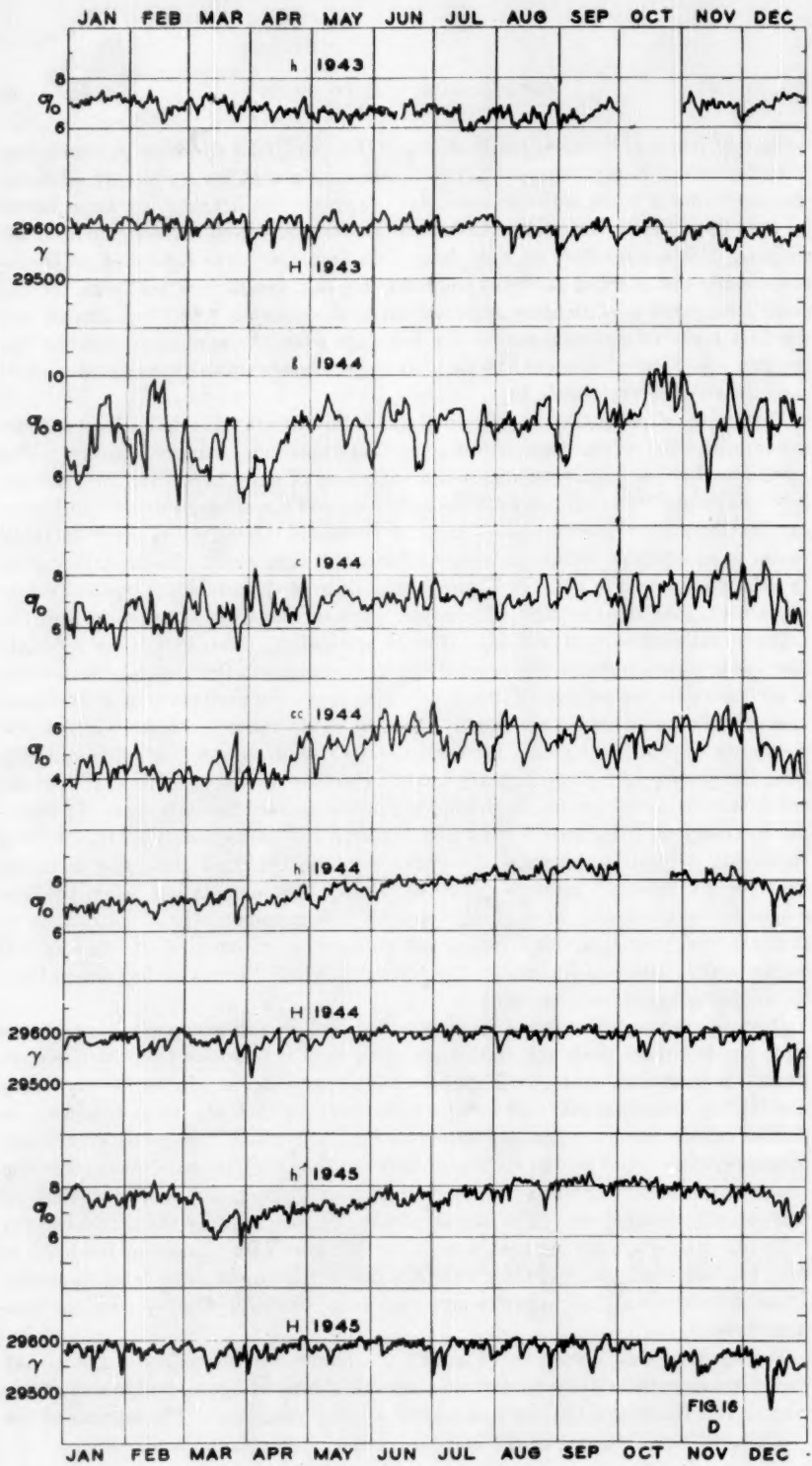


FIG. 4.—Daily means, cosmic-ray intensity, 1944 (see Fig. 3).

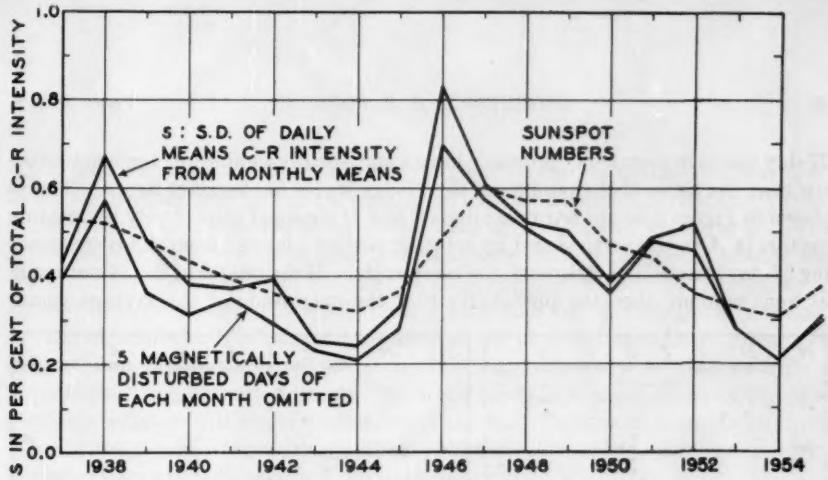


FIG. 5.—Annual means: sunspot numbers and variability, cosmic-ray intensity at Huancayo.

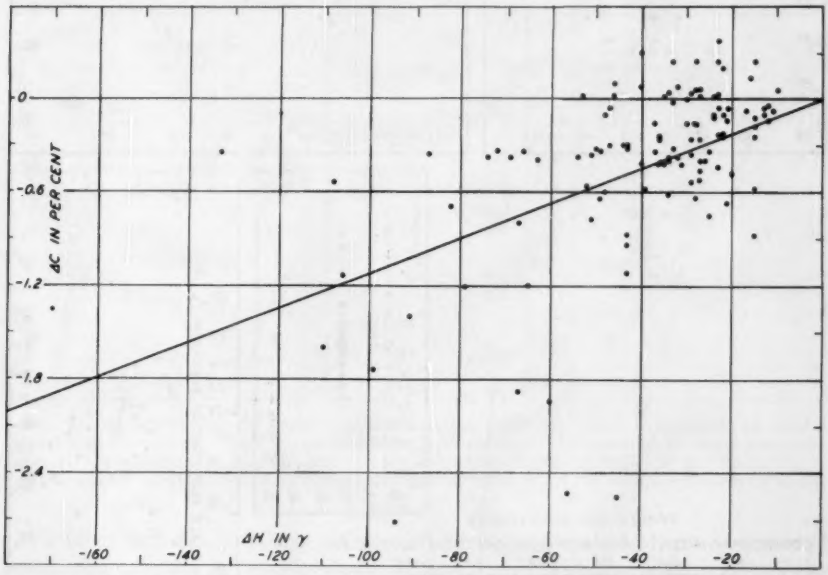


FIG. 6.—Average difference for cosmic-ray intensity (ΔC) and for horizontal magnetic field (ΔH) for five disturbed days less that for five quiet days in each month, April, 1937–December, 1946, at Huancayo.

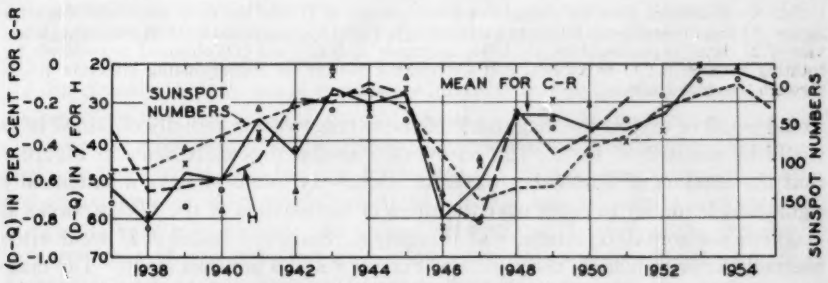


FIG. 7.—Annual means for magnetically disturbed less quiet days (5 each per month) for cosmic-ray intensity (C-R) and for horizontal magnetic intensity, H.
Legend for C-R: ° = Cheltenham, x = Huancayo, Δ = Godhavn.

27-day waves in cosmic-ray intensity have a statistically significant[2] tendency to occur near the times of the minima of the 27-day waves in character figure. This is shown in Figure 8 by the harmonic dials C and D obtained respectively by rotating vectors in A to the vertical and by rotating vectors (derived from the corresponding 27-day interval) in B through the same angle. If the phases of the 34 vectors in D were random, then the probability that the magnitude of the average vector

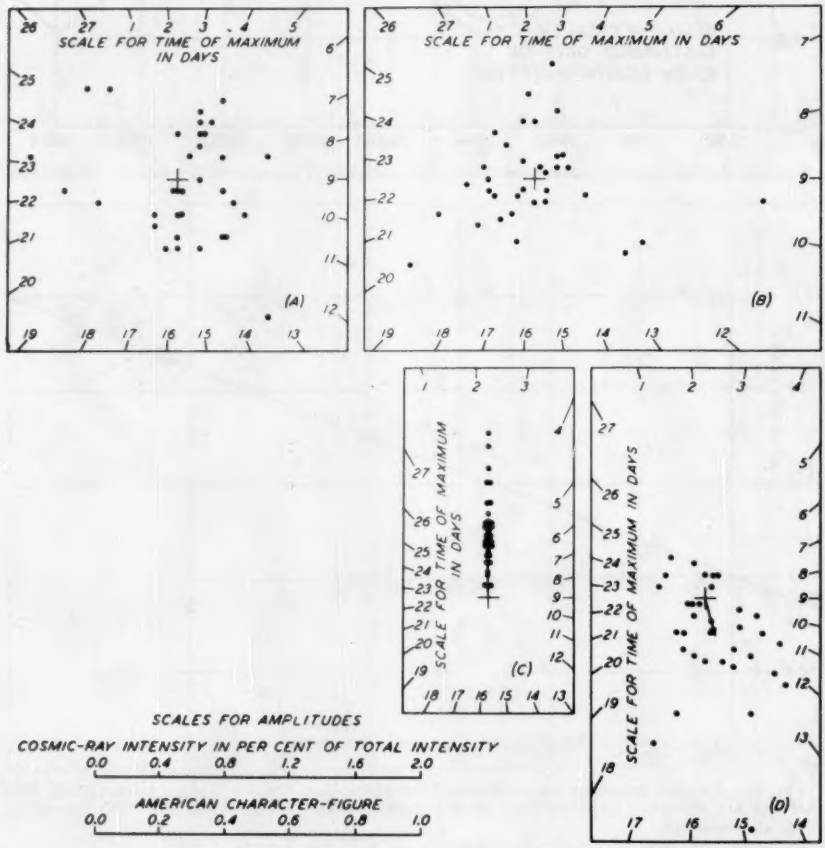

SCALE FOR TIME OF MAXIMUM IN DAYS
(A)
SCALE FOR TIME OF MAXIMUM IN DAYS
(B)
SCALE FOR TIME OF MAXIMUM IN DAYS
(C)
SCALE FOR TIME OF MAXIMUM IN DAYS
(D)
SCALES FOR AMPLITUDES
COSMIC-RAY INTENSITY IN PER CENT OF TOTAL INTENSITY
0.0 0.4 0.8 1.2 1.6 2.0
AMERICAN CHARACTER-FIGURE
0.0 0.2 0.4 0.6 0.8 1.0

FIG. 8.—Harmonic dials for departures from average of 27-day waves in American character figure (A) and in cosmic-ray intensity at Huancayo, Peru (B), computed from 34 rotations (intervals of 27 days) beginning June 13, 1936; harmonic dials (C) and (D) obtained respectively by rotating vectors in (A) to vertical and by rotating vectors for corresponding intervals in (B) through the same angle.

would equal or exceed the magnitude of the average vector actually obtained in D would be about 2×10^{-6}. The results of a similar procedure show in Figure 9 that the maxima of the 27-day waves in cosmic-ray intensity have a statistically significant tendency to occur near the times of the maxima of the 27-day waves in magnetic horizontal, H, intensity at Huancayo. Since low values of H occur when magnetic activity is high, the results of Figures 8 and 9 are consistent. The magnetic-storm effects, the (magnetically) disturbed minus quiet-day differences, and

the 27-day variations in cosmic-ray intensity and in the horizontal component, H, of the earth's magnetic field at the equator all indicate a significant tendency for decreased values of cosmic-ray intensity to occur with decreased values of H. The latter decreases are known to arise from magnetic fields with sources outside the earth. Thus the mechanism for most of these changes in cosmic-ray intensity is closely connected with the mechanism responsible for the magnetic changes. Several attempts have been made to calculate the expected cosmic-ray changes in magnetic storms, assuming that the external magnetic-storm field arises from a hypothetical ring current concentric with the earth and in the plane of the geomagnetic equator (this would explain one of the main features of magnetic storms). None of the results obtained has indicated that the decreases in cosmic-ray intensity arise from magnetic effects of the ring current.

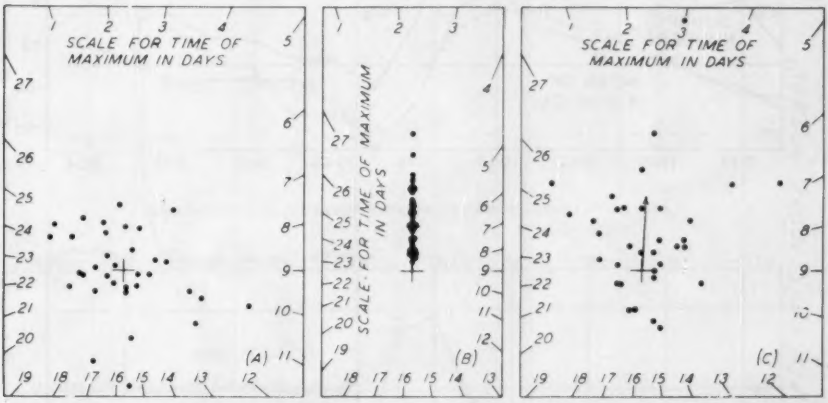


FIG. 9.—Harmonic dials for departures from average of 27-day waves in magnetic horizontal intensity at Huancayo, Peru (A), computed from 34 rotations (intervals of 27 days) beginning June 13, 1936; harmonic dials (B) and (C) obtained respectively by rotating vectors in (A) to vertical and by rotating vectors for corresponding interval in (B) of Fig. 8 through the same angle.

Variation with Sunspot Cycle.—Figure 10 shows the variation of annual means of cosmic-ray intensity at four stations and a comparison with sunspot numbers of the annual means averaged for all stations for the period 1937–1955. The latter curve indicates that cosmic-ray intensity is higher near the sunspot minima. This 11-year variation in cosmic-ray intensity at Huancayo, as shown in Figure 11, is about the same for all days, for magnetically disturbed days (5 per month), and for magnetically quiet days (5 per month). This indicates that the 11-year variation is not due directly to decreases in cosmic-ray intensity during magnetic storms, which are more frequent near times of sunspot maxima. It has been suggested that conducting solar streams (which give rise to magnetic storms) which carry "frozen-in" magnetic fields away from the sun during sunspot maxima may pervade the solar system to an extent which would reduce the flux of cosmic rays arriving at the earth from outside the solar system. In connection with the 11-year variation, it should be noted that Meyer and Simpson,[6] using nucleonic detectors in jet aircraft at 30,000 feet altitude, found that the knee of their latitude curve moved northward 3° between 1948 and 1951, showing that at the latter time additional low-rigidity

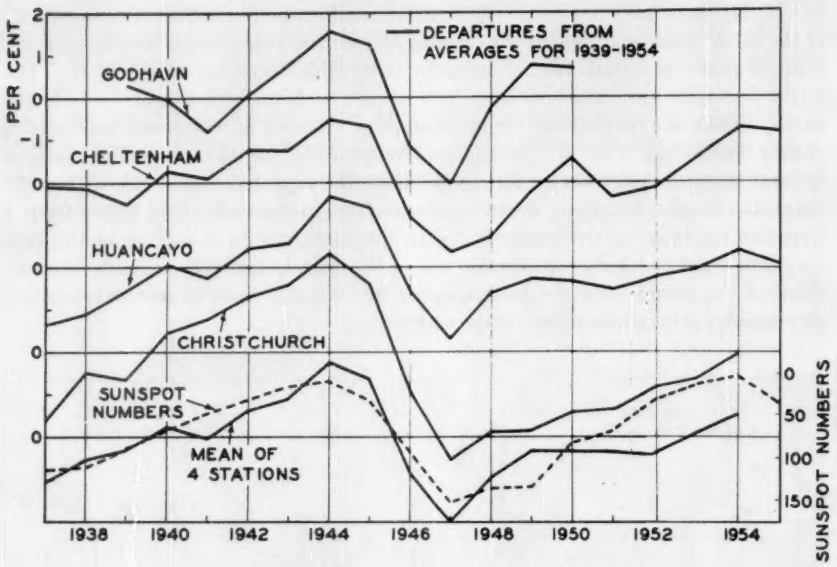


FIG. 10.—Annual means, cosmic-ray intensity, at four stations.

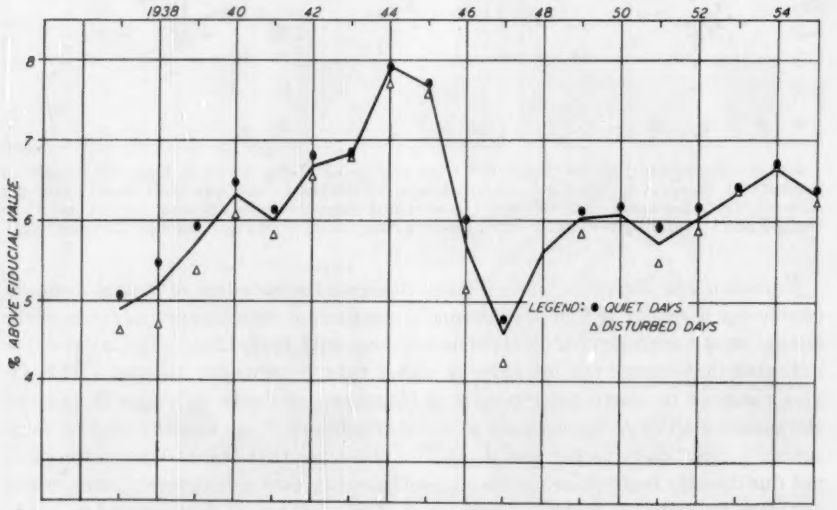


FIG. 11.—Annual means, cosmic-ray intensity, at Huancayo for all days, international magnetic quiet days, and international disturbed days.

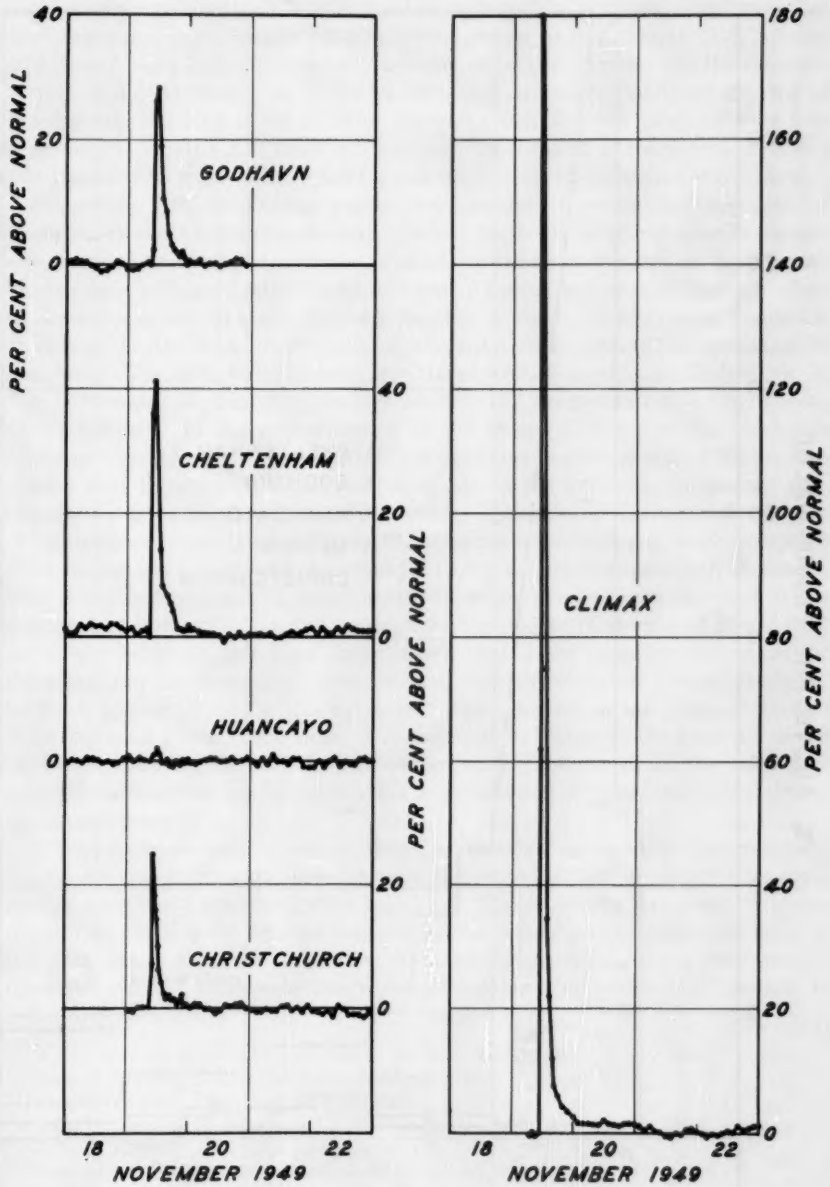


FIG. 12.—Increase of cosmic-ray intensity, November 19, 1949.

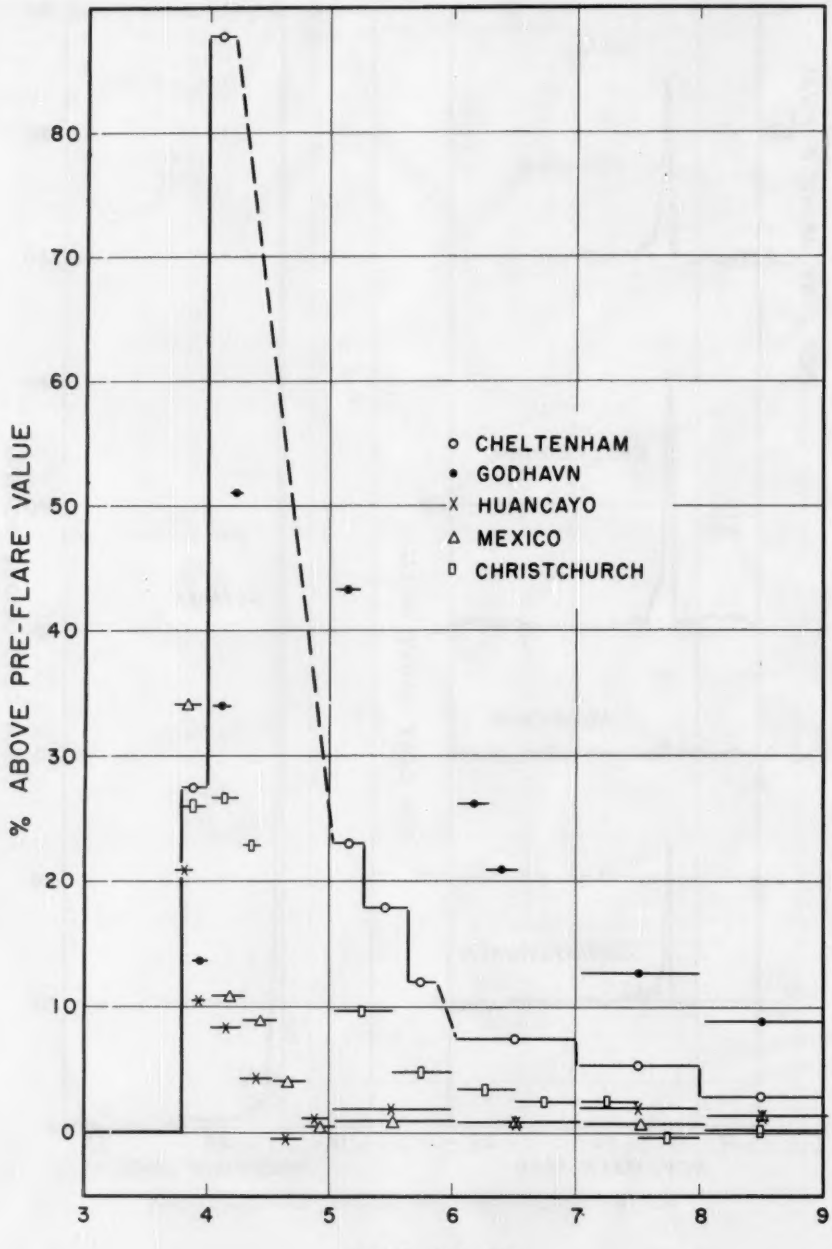


FIG. 13.—Cosmic-ray intensity following solar flare at 0330 g.m.t., Feb. 23, 1956.

particles were reaching this level in the atmosphere. In addition, they found that the total cosmic-ray intensity was 13 per cent greater in 1951 than in 1948. Neher and Stern,[7] using high-altitude balloon-borne ionization chambers in the summers of 1951 and 1954, found, at the latter time, new low-energy particles arriving at geomagnetic latitudes north of 58°. Assuming that the new particles were protons, they calculated that there was no magnetic cutoff of primary protons above 150 Mev in 1954, whereas the cutoff for protons in 1951 was estimated at 800 Mev.

Solar-Flare Effects.—During nearly two decades of continuous operation of Compton-Bennett ionization chambers, five[4, 5] unusually large increases in cosmic-ray intensity have been recorded. Each of these followed within an hour (in all but one case, within a quarter-hour) the onset of a solar flare or radio fadeout. One of these increases is shown in Figure 3, another in Figure 12, and one in Figure 13. In four of the five increases no increase occurred at Huancayo (at the geomagnetic equator). The extra large increase at Climax relative to that at Cheltenham in Figure 12 was due mainly[4] to the high altitude of Climax relative to Cheltenham (3,428 meters). In the second section of this review ("Origin of the Nucleonic Component and the Solar-Flare Effect") the arguments were presented for the conclusion that the solar-flare increases were due to the nucleonic component generated by a great increase in the flux of charged particles in the lower-energy part of the cosmic-ray spectrum. Figure 13 shows the recent increase of February 23, 1956, as measured at several stations with Compton-Bennett ionization chambers. This is the first occasion on which a solar-flare increase has been observed at the equator and shows that charged particles with momenta of at least 15 Bev/c were involved. Schlüter[8] and Firor[9] have shown that if the particles responsible for the solar-flare increases come from the sun, then the observed intensities should be much greater in certain "impact zones" which depend on the geomagnetic latitude and local geomagnetic time. For example, in Figure 13 the large difference in the intensity at Cheltenham and Christchurch (both at about the same geomagnetic latitude, except for sign) is due to the location of these stations relative to the impact zones.

It thus seems reasonably certain that the solar-flare increases in cosmic-ray intensity are due to charged particles from the sun which are accelerated by some mechanism closely associated with the flare. The remaining variations of intensity are doubtless perturbations imposed by the magneto-hydrodynamical state of the solar system upon a steady influx of cosmic-ray particles coming from outside the solar system. The mechanism which effects these perturbations of intensity is not clearly understood.

[1] A. H. Compton, E. O. Wollan, and R. D. Bennett, *Rev. Sci. Instr.*, **5**, 415, 1934.
[2] Scott E. Forbush, *J. Geophys. Research*, **59**, 525, 1954.
[3] Scott E. Forbush, *Phys. Rev.*, **54**, 975, 1938.
[4] Scott E. Forbush, M. Schein, and T. B. Stinchcomb, *Phys. Rev.*, **79**, 501, 1950.
[5] Scott E. Forbush, *J. Geophys. Research*, **61**, 155, 1956.
[6] Peter Meyer and J. A. Simpson, *Phys. Rev.*, **99**, 1517, 1955.
[7] H. V. Neher and E. A. Stern, *Phys. Rev.*, **98**, 845, 1955.
[8] A. Schlüter, *Z. Naturforsch.*, **6a**, 613, 1951.
[9] J. Firor, *Phys. Rev.*, **94**, 1017, 1954.

*SOLAR ORIGIN OF CHANGES IN THE PRIMARY COSMIC RADIATION**

By J. A. Simpson

ENRICO FERMI INSTITUTE FOR NUCLEAR STUDIES AND DEPARTMENT OF PHYSICS, UNIVERSITY OF CHICAGO

I. Introduction.—Until a few years ago, a strong impression persisted that most observable intensity variations were probably of terrestrial origin, being due to variations of meteorological effects or variations of our earth's magnetic field, and that any relationship of these variations to solar phenomena arose through indirect processes occurring on the earth. It has now become increasingly clear from experiments and studies over the past six to eight years, however, that many of these variations of cosmic-ray intensity are in fact variations of primary particle intensity in the vicinity of our earth—how close to our earth we do not yet know. During these past few years, it has also become firmly established that the variations represent changes in the shape of the primary differential particle spectrum distribution. More specifically, it is the low-energy portion of the cosmic-ray spectrum which undergoes the largest intensity changes with time.

Several of these new observations have been correlated with properties of the sun, and it is now abundantly clear that the principal changes in the primary spectrum are associated with solar phenomena—broadly speaking, the intensity variations have a solar origin. It is only at the very high energies—energies much higher than we shall be concerned with here—that primary variations are thought to depend on sidereal time.

Indeed, the search for a theory of the solar origin of these intensity variations is leading to a new description of the electromagnetics of the interplanetary medium and the nature of solar active regions which may have astrophysical and geophysical implications extending beyond the explanation of cosmic-ray intensity variations and the origin of cosmic radiation.

These developments have come about partly as a result of extending observations into the low-energy end of the primary spectrum, while simultaneously continuing the excellent observational work begun over the past twenty or more years at higher energies with ion chambers and counter telescopes.[1] This point is illustrated in Figure 1, where the primary differential flux j is shown as a function of the primary particle rigidity. The rigidity of a particle of charge Ze and momentum P in a magnetic field is given by Pc/Ze—all particles of the same rigidity will experience the same deflection in a magnetic field. [Rigidity is measured in billion volts.] From our knowledge of the detector response to secondary particles generated by particles in this primary spectrum, we can say roughly what portion of the primary spectrum is being observed within our atmosphere by any given detector. In Figure 1 we see that a shielded ion chamber or meson telescope is capable of extending measurements from approximately 7–10 BV to the highest rigidities. The nucleonic component, as measured by local neutron production, extends from approximately 1 BV to the higher rigidities and thus responds to over 95 per cent of the particles arriving in the primary beam. For even lower rigidities, one must resort to balloon and aircraft-borne instruments. As the primary spectrum undergoes a change of total intensity (say, between times t_1 and t_2),

observations from all these instruments, taken together with the geomagnetic field as a spectrum analyzer, provide us with means for dividing the spectrum into regions of limited range of particle energy or rigidity and enable us to learn about the rigidity or energy dependence of the variation with time. It is this coverage of the spectrum which gives us the possibility for analyzing the variations in new ways.

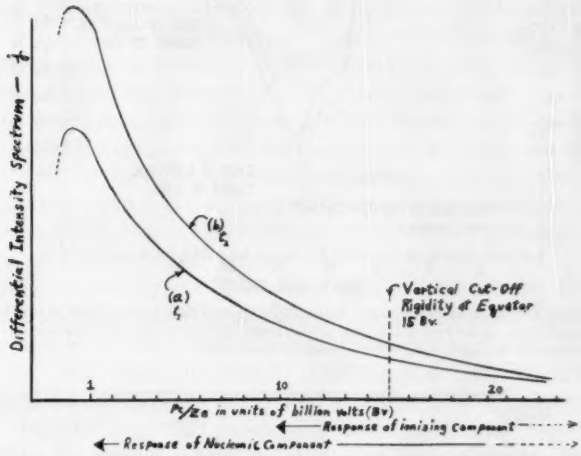


FIG. 1. Intensity of primary cosmic radiation as a function of incident particle momentum. When the cosmic radiation undergoes a change of intensity between times t_1 and t_2, the spectrum changes from a to b.

In fact, by these techniques we demonstrate that neither meteorological fluctuations nor variations of the earth's magnetic field are the origin of the cosmic-ray variations discussed here. In Figure 2 we show a small, but typical, Forbush-type decrease of intensity measured simultaneously by neutron detector and ionization chamber.[1] Note that the percentage change in amplitude of the neutron observations is about four times that for the shielded ion chamber of Freiburg, Germany. Thus one concludes that the amplitude of the variation is indeed dependent upon the rigidity of the primary particles. If this world-wide decrease of intensity were due to changes of the earth's field, the integrated intensity above the knee of the latitude curve for the neutron monitor would show no appreciable change of intensity. To test this, observations were made near the time of minimum intensity and again near the time of normal intensity, as shown in Figure 2, with the results shown in Figure 3.[2] These data were taken at approximately 300 gm/cm^2 atmospheric depth. We can immediately see how this change in integrated rate may come about, by writing the following equation:

$$\text{Count rate } (\lambda, x, t) \propto \int_{N_0}^{\infty} S(x, N) j(N, t)\, dN,$$

where λ is magnetic latitude, x is atmospheric depth, t is the time, $N = Pc/Ze$, N_0 is the vertical geomagnetic cutoff rigidity as determined from Störmer theory, and S is the function relating the generation of secondary particles to the primary

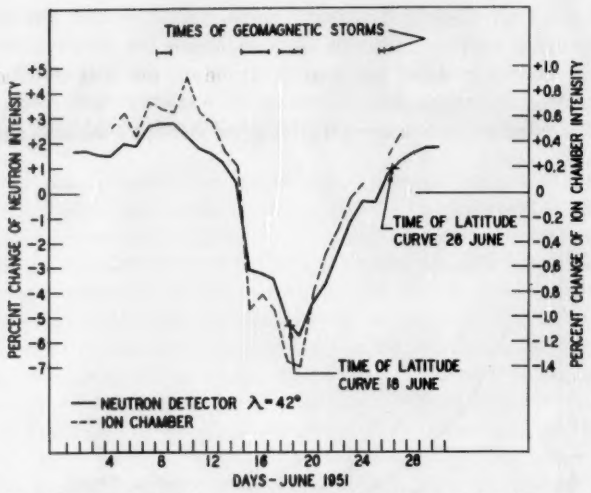

TIMES OF GEOMAGNETIC STORMS
PERCENT CHANGE OF NEUTRON INTENSITY
PERCENT CHANGE OF ION CHAMBER INTENSITY
TIME OF LATITUDE CURVE 26 JUNE
TIME OF LATITUDE CURVE 18 JUNE
NEUTRON DETECTOR λ=42°
ION CHAMBER
DAYS - JUNE 1951

FIG. 2. A sharp decrease of cosmic-ray intensity arising from a modulation mechanism of solar origin. The amplitude observed with the low-energy neutron detector is about a factor of 5 times the amplitude in a high-energy ion-chamber detector.

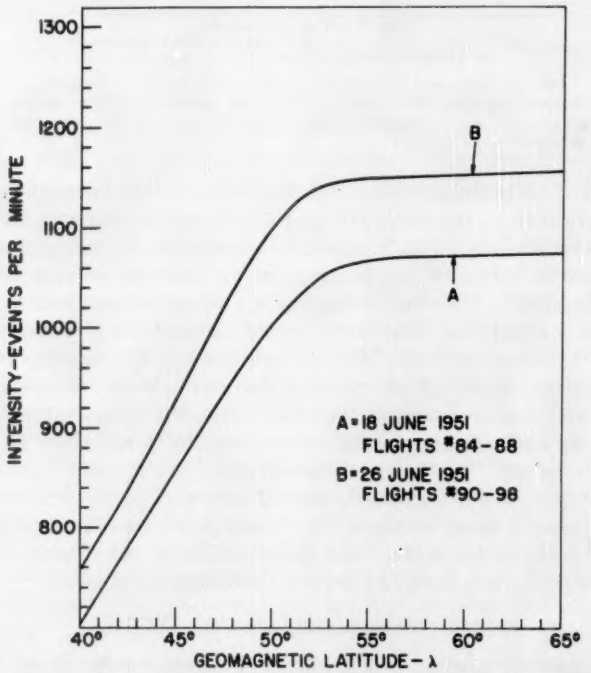


FIG. 3. The change of the secondary component intensity arising from primary radiation during the modulation effect shown in Fig. 2. This shows that the modulation is not due to the geomagnetic field.

radiation. We see that the variation in j produces a variation of intensity in the counting rate. There are other arguments which, combined with the above, make the concept of axially symmetric variations of the geomagnetic field (for example, the influence of an external ring current) unacceptable for explaining the origin of the cosmic-ray variations.

It has also been shown by careful series of experiments that the principal intensity variations are not produced by atmospheric changes. Hence we are led to consider the extraterrestrial origin of the cosmic-ray variations which are most apparent at particle rigidities below approximately 60 BV rigidity—which rigidity range is defined as the low-energy, or low-rigidity, primary spectrum, containing over 95 per cent of all the particles in the cosmic radiations. From our present knowledge of the physical properties of intensity variations, all variations can be classified as either the occasional, *de novo* production of cosmic-ray particles in the vicinity of the sun or the modulation of pre-existent cosmic rays, presumably from outside our solar system. Examples of this latter type of variation, all of which are nonperiodic, are the 27-day recurrences and Forbush-type decreases of cosmic-ray intensity.

II. *The Solar-Flare Effect.*—There is a unique phenomenon in cosmic rays whereby we are able to observe optically the onset of a solar mechanism which leads to the arrival of additional cosmic-ray particles at the earth. This is the solar-flare effect, discovered in 1942 by Forbush and Ehmert. Since these are temporary increases of intensity many fold above normal, it is indeed difficult to invoke any origin for this effect except production near or on the sun. We shall not outline here the details or features of the five large intensity increases, or the less well-established small flare effect. However, from either the large or the small flare effect, there is a tendency for particles to arrive in special impact zones or bands of longitudes, at intermediate latitudes, with the description of the impact zones being given by the discrete Störmer orbits for particles traveling between the sun and the earth. Schlüter and Firor[3] have investigated these zones, and the latter has considered the distribution of the large flares within the framework of this simple hypothesis. In Figure 4 we see the range of impact zones for the region of highest intensity (approximately 0900 hour) and for other zones of lower intensity. Actually, these zones are spread out in time by approximately four hours. These results lead to the elementary conclusion that, if a detector is within an allowed zone at the time of a cosmic-ray-producing flare, an increase of intensity will be observed.

However, it is well established that increases are also observed at very high magnetic latitudes for which trajectories between the sun and the earth are not at present known to exist—and on a few occasions there are perhaps increases outside the predicted impact zones.

These difficulties emphasize that the simplest views regarding impact zones require further revision and bring into sharp focus the problem of describing the behavior of solar-produced cosmic-ray particles in interplanetary space.

Of all the solar-flare events, the cosmic-ray production associated with the February 23, 1956, chromospheric eruption was the largest and most completely studied;[4] hence we shall use it for our example. The change of intensity observed at Chicago is shown in Figure 5. Let us examine some of the experimental facts:

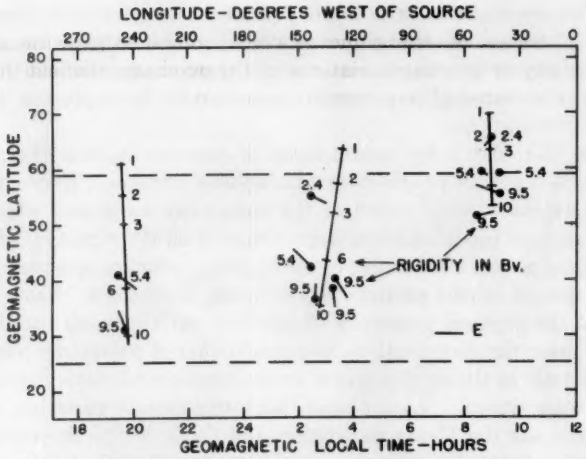

LONGITUDE - DEGREES WEST OF SOURCE
270 240 210 180 150 120 90 60 30 0
GEOMAGNETIC LATITUDE
80 70 60 50 40 30 20
18 20 22 24 2 4 6 8 10 12
GEOMAGNETIC LOCAL TIME - HOURS
RIGIDITY IN Bv.
E

FIG. 4. Impact regions for cosmic-ray orbits connecting the sun with the earth when the sun lies in the plane of the magnetic equator.

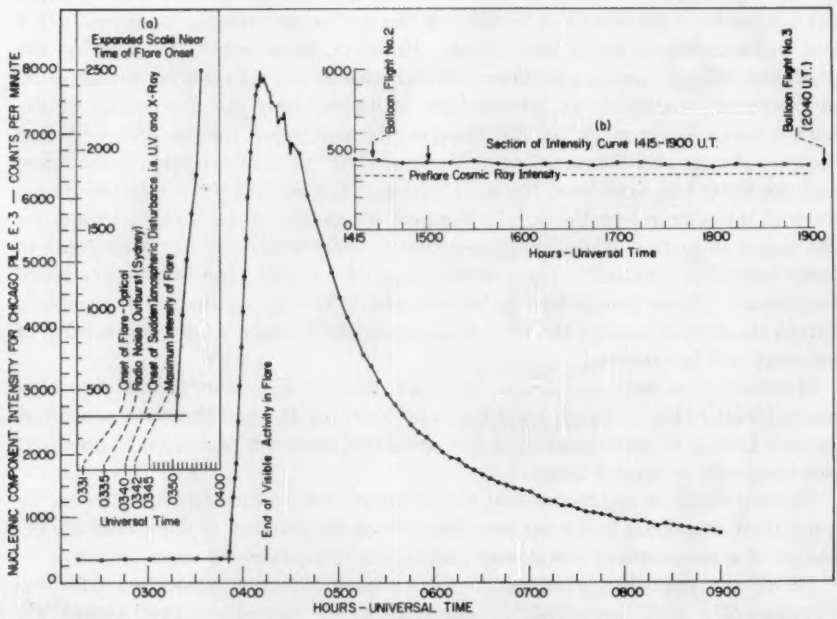

NUCLEONIC COMPONENT INTENSITY FOR CHICAGO PILE E-3 — COUNTS PER MINUTE
(a)
Expanded Scale Near Time of Flare Onset
Onset of Flare - Optical
Radio Noise Outburst (Sydney)
Onset of Sudden Ionospheric Disturbance i.e. UV and X-Ray
Maximum Intensity of Flare
Universal Time
End of Visible Activity in Flare
HOURS - UNIVERSAL TIME
(b)
Section of Intensity Curve 1415-1900 U.T.
Balloon Flight No.2
Balloon Flight No.3 (2040 U.T.)
Preflare Cosmic Ray Intensity
Hours - Universal Time

FIG. 5. The temporary increase of cosmic-ray intensity observed with the neutron pile at Chicago, February 23, 1956. The particles producing this increase of intensity were accelerated in association with a large solar flare.

a) A typical flare effect observed at intermediate latitudes is represented by a rapid rise of intensity (the pulse front), followed by a slow decline of intensity persisting for a period of many hours (the tail). Most likely, this rapid increase is due to the almost direct arrival of particles from the vicinity of the sun, whereas interpretation of the tail is much more complex, if for no other reason than that it persists for many hours beyond the expected lifetime of the solar event. Due to uncertainties, especially in the recording methods of different observers, we are not yet certain whether the rate of rise of the intensity is different for different flares. Since, for at least some of the flare events, the major part of the intensity rise occurs in an interval of a few minutes, it is most probable that the particles have come from the sun along relatively uncomplicated orbits and, hence, have undergone relatively little scattering.

b) For the tail of the flare increase we examine the rate of decay as a function of incoming particle energy. The data from the neutron intensity monitors (Fig. 6)[2] extending northward from the geomagnetic equator are shown in Figure 7. It is obvious that the high-energy particles disappear more rapidly from the incoming flux of flare particles than the low-energy particles—the flare spectrum is time dependent. From these data the spectrum has been reconstructed for three different times following the onset of the flare (Fig. 8), from which we note that the flare spectrum after several hours contains only low-energy particles. We have confirmed this conclusion by high-altitude balloon flights during the progress of the flare.[4] There are two alternative hypotheses which would appear to account for these results: (1) The interplanetary medium may contain a large number of scattering and diffusing centers produced by random magnetic fields, so that after the initial burst of radiation from the sun, a portion comes almost directly to the earth to produce a rapid rise of intensity. The remainder scatter within the solar system and arrive subsequently from *all* directions in the sky. Consequently, the highest-energy particles will diffuse out of the solar system faster than low-energy particles, so as to produce an energy dependence for the decay rate of the solar-flare effect. Alternately (2) direct production in the region of the sun continues many hours after the optical and radio effects have ceased.

Neither the second alternative nor proposals that the flare particles are trapped in special orbits in the solar system appear to provide a satisfactory physical interpretation of the observations.

c) Since a large increase of intensity was observed at the geomagnetic equator,[4] the flare-production spectrum contained particles in excess of 15 BV rigidity and most likely an appreciable flux of more than 20–30-BV particles. From the balloon flight we know that the rigidity spectrum extends below 1.5 BV. Our knowledge of the spectrum near the time of onset is uncertain, since we do not yet understand in detail the dependence of particle intensity upon the longitudinal position of the detector at the time of the event.

d) We may estimate the maximum angle of particle emission from the region of solar flares by examining the distribution of co-ordinates on the sun of those flares which produce the large and small flare effects. From Figure 9 it is clear that emission over a large solid angle occurs—the November 19, 1949, flare being at 70° west heliolongitude. However, only the two large flare effects which were near central meridian (the location of the event for March, 1942, was not directly

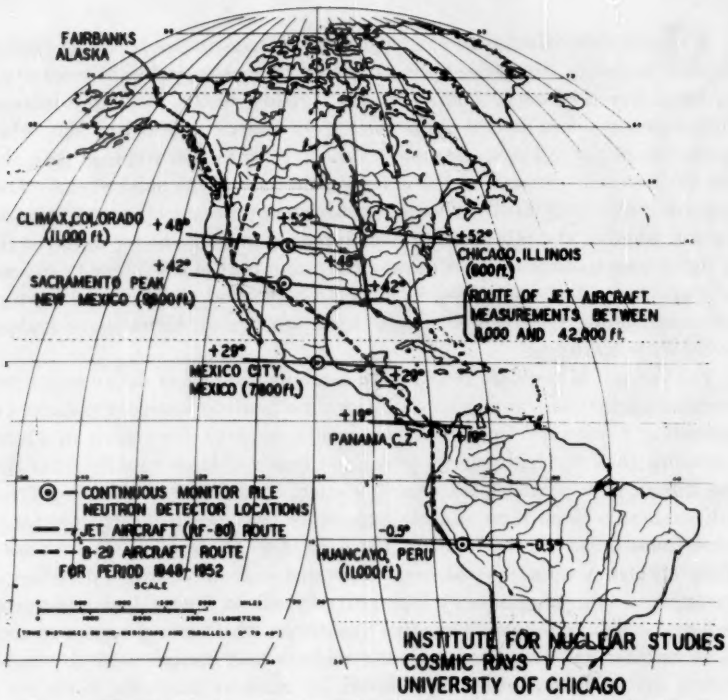


FIG. 6. The distribution of neutron piles which observe continuously the changes of cosmic-ray intensity. The geomagnetic field acts as a momentum analyzer for this network of detectors.

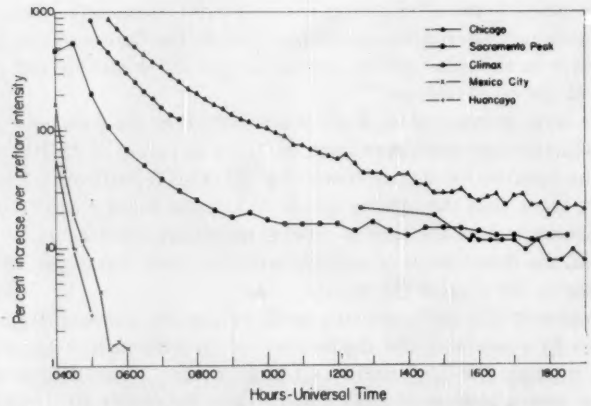


FIG. 7. The decline of cosmic-ray intensity following the temporary increase of February 23 at the neutron monitors shown in Fig. 6.

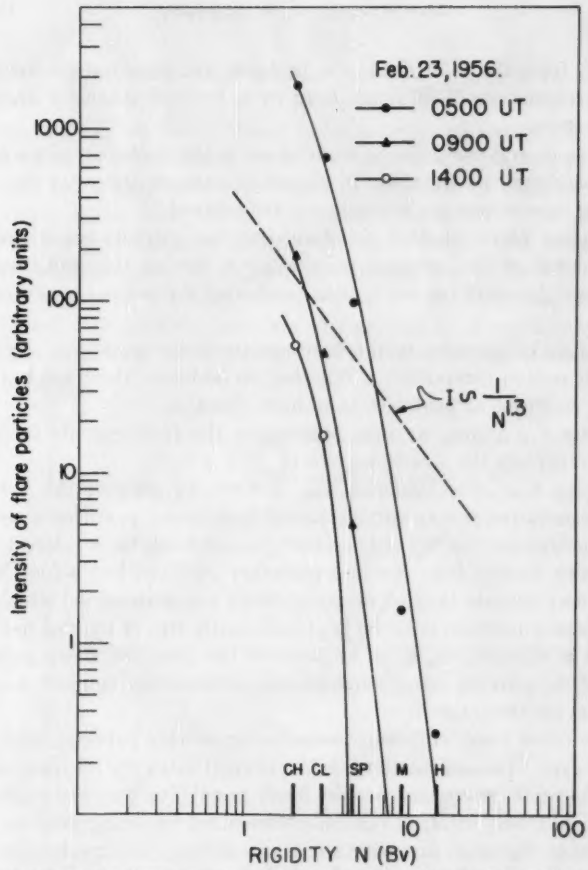


FIG. 8. The spectrum of primary cosmic-ray particles produced by the flare derived from the data in Fig. 7. The normal cosmic-ray spectrum is shown as a dashed line, but with arbitrary intensity. $N = pc/ze$.

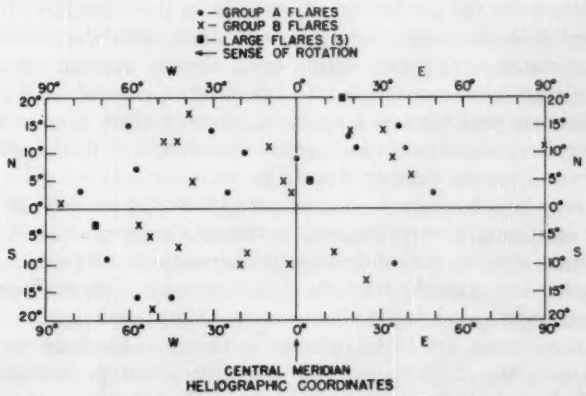


FIG. 9. The locations of solar flares on the solar disk which produced increases of cosmic-ray intensity.

observed but, from the solar data, was probably far from central meridian) were followed approximately 25–30 hours later by a Forbush intensity decrease and a geomagnetic storm.

e) Since we do not yet know at what phase in the evolution of a solar flare the associated cosmic-ray particles are produced, the transit times for the particles in the four large flare events are not uniquely determined.

f) We do not know whether the flare spectrum extends below the prevailing low-rigidity cutoff of the isotropic radiation. A test on this point would be important for any theory of the mechanism producing the low-energy cutoff of cosmic rays.

g) The kinds of particles in the flare spectrum are unknown. Evidence supports a major proton component. Whether, in addition, there are heavier nuclei, and possibly electrons or neutrons, is an open question.

From points *a–g* above, we may summarize the requirements for any theory attempting to explain the solar-flare effect:

1. Assuming that new evidence will continue to support the view that the process represents temporary production of cosmic-ray particles associated with the flare, a mechanism for the acceleration process must be developed.

2. It is now known that the interplanetary medium has a high average ion density and may contain tangled magnetic fields and perhaps ion streams. Hence the interplanetary medium must be kept sufficiently free of tangled fields or beams at the times of flare events, so as to preserve the observed sharp pulse fronts of particles and to provide some longitudinal asymmetries (impact zones) at the earth (at least for the pulse front).

3. On the other hand, sufficient means for scattering particles within the solar system must exist to obtain an appreciable incident intensity for many hours in the vicinity of the earth, or require the less likely possibility that the solar production process continues long after all optical evidence for the solar process has ceased.

After we have discussed intensity variations arising from modulation processes, we shall again return to this problem and indicate how theories for both kinds of phenomena are linked together.

III. *Modulation Mechanisms Acting on the Primary Radiation.*—Whereas there is strong evidence for the production of particles in the solar-flare effect, the remaining nonperiodic intensity variations with time have characteristics which suggest a modulation mechanism acting upon already existing cosmic-ray particles. Modulation mechanisms fall into two distinct classes:

1. Particles may pass through a region in space in which they receive a small acceleration or deceleration, and the integral spectrum will then appear to have either a larger or a smaller number of particles, or

2. There may be a temporary removal (or addition) of particles in the cosmic-ray beam by scattering, absorption, etc., so that at a given position in space there will appear from time to time decreases (or increases, if albedo effects exist) of integrated cosmic-ray intensity from the mean intensity. We shall consider later the evidence supporting either of these classes of modulation mechanisms. Intensity variations which are to be included in this discussion are the 27-day recurring variations, the sharp decreases of intensity (Forbush decreases), the 11-year cycle of changing cosmic-ray intensity, and the changes in the low-rigidity cutoff of the spectrum which have recently been shown to exist.

There are, first of all, some observations that refer generally to all these intensity changes:

a) The amplitude of all these intensity changes is dependent on the magnetic rigidity of particles being measured, with the amplitude being largest for lowest rigidity but still being discernible for magnetic rigidities as high as 30–40 BV. The rigidity dependence, with changes in intensity, appears to be of a slightly higher power dependence than the primary spectrum itself. For example, if the spectrum has the form *a* given in Figure 1, at time t_1, then at time t_2 an intensity increase would appear as a change to spectrum *b*. By including low particle rigidities measured with neutron detectors, the amplitude of the 27-day variation may occasionally be as large as 20 per cent. The Forbush-type decreases undergo even larger changes.

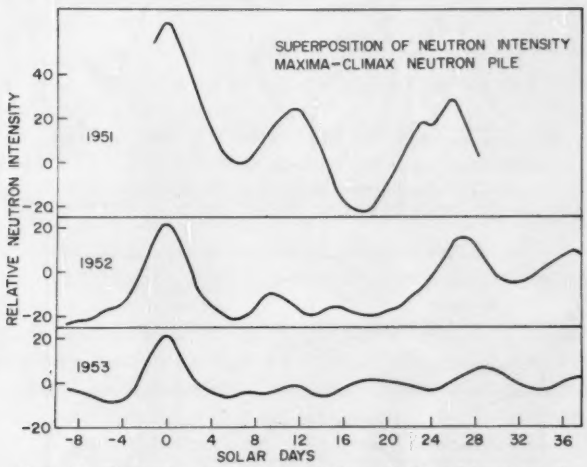


FIG. 10. Recurring intensity variations of the primary cosmic-ray intensity having an approximate recurrence of 27–28 days.

What has earlier been assumed to be a time-invariant low-energy cutoff in the primary spectrum has now been found to change with time in such a way that no lower limit for particle energies is found at solar minimum, whereas a 1-Bev lower limit (or maximum in the differential intensity curve) is found at maximum solar activity. This change in the low-energy spectrum is also energy dependent.

b) These intensity variations are all associated with the magnetic solar cycle on the sun, there apparently being a correlation over the solar cycle with the amplitude of the 27-day variations (Figs. 10 and 11).[4] The excellent long-time recordings made by Forbush give the evidence for a 11-year solar-cycle-dependent intensity variation[1] (Fig. 12). Although efforts to correlate solar phenomena with cosmic-ray intensity variations leave much to be desired, it is already evident that modulation mechanisms must have a solar origin.

c) The intensity variations all appear to possess the property of maintaining isotropy in the region of the earth. This is a very important observation which places severe restrictions on any theory for the modulation process.

d) Clearly, the choice of a modulation mechanism will depend on whether there occur only decreases of intensity from the mean cosmic-ray intensity or whether there are both increases and decreases of intensity from the mean. The Forbush-

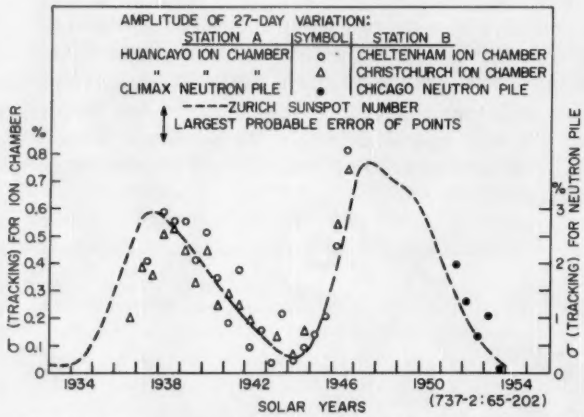


FIG. 11. The average amplitude of the 27–28-day intensity variation changes from year to year following the ~ 11-year solar cycle.

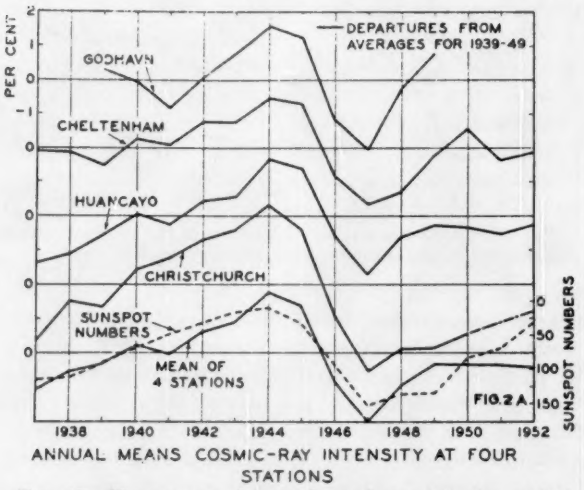


FIG. 12. The average change of cosmic-ray intensity at high energy over an entire solar cycle. Found by Forbush, using ionization chambers.

type sharp decrease is clearly an example of an effect which argues for only decreases of intensity from the mean value. However, it does not seem to be a simple matter to argue that the 27-day variation (and 24-hour variations, which will not be considered here) are only decreases of intensity from the mean value.

There are many other facts known about these nonperiodic variations. The general properties selected here represent the kinds of information we have upon which to base any theory for the solar origin of time variations. Over the past few years there have been two independent hypotheses, one resting upon rather well-defined ion streams issuing from the sun and carrying a "frozen-in" magnetic field along with the stream. This has been proposed by Alfvén and is well known. The other concept rests upon the emission of expanding cloudlike regions from the sun containing twisted or tangled magnetic fields intermixed with ionized clouds of matter having velocities of the order of 10^3 km/sec. This concept of magnetized clouds has been suggested by Fermi, Gold, Schlüter, Morrison,[5] and others and requires that the first-order effects on the cosmic-ray intensity are decreases from the undisturbed extrasolar intensity. In contrast, the Alfvén beam requires that acceleration or deceleration of particles be the major process. Both concepts lead to a form of energy or rigidity dependence for the amplitude for the intensity variations.

Both concepts also lead to the prediction of geomagnetic storm effects and correlations with solar active regions as required by any theory capable of explaining the cosmic-ray intensity variations.

Although neither concept is at present fully developed and capable of explaining all the observed facts concerned with cosmic-ray changes or geomagnetic storms, it is apparent that the diffusion of cosmic rays through magnetic clouds, as proposed by Morrison, and more recently in a different theory by Parker,[6] appears to be the more satisfactory qualitative model, particularly since it is a statistical process, leads to a high degree of isotropy near the earth, and may account for some features of the geomagnetic storm.

Let us consider how diffusion of cosmic-ray particles through ionized clouds containing twisted and knotted fields may introduce changes of total cosmic-ray intensity. For purposes of illustration, we assume that the diffusing region is of infinite extent and thickness y and that the full galactic intensity is observed in region *1* (Fig. 13, *a*). Then a detector placed in region *2* after the system has reached equilibrium will detect the full galactic radiation intensity. If, however, region *2* is initially free of radiation, and the galactic radiation begins to diffuse through the scattering barrier, the detector will observe that (1) the intensity rises asymptotically to the galactic intensity and (2) the rigidity dependence of the spectrum is changing and approaches the spectrum of the galaxy.

The flux of particles $F(x, y, z; E)$ in region *2* arising from a gradient in the cosmic-ray particle density $J(x, y, z; E)$ for the mean free path $L(x, y, z; E)$ of magnetic scattering centers ($L \ll y$) in the twisted-fields region may be written

$$F(x, y, z; E) = -\frac{L}{3}\nabla[vJ]$$

for particles of velocity v.

This is the mechanism proposed by Morrison to account for the intensity variations. A magnetized cloud of ionized matter is ejected from the sun, expanding to an enormous volume which initially contains no cosmic radiation in its inner regions (Fig. 13, *b*). If, then, the earth sweeps into this volume, a detector will measure a sharply reduced intensity. The outward-moving cloud also "sweeps

out" preferentially the low-energy particles due to its velocity v'. The time to reduce the total intensity for reasonable field intensities is of the order of a day or more, with the time for recovery to normal intensity of the order of days.

On the other hand, if we introduce an *absorber* in region *2* which is capable of removing particles as fast as they diffuse through the barrier, as in Figure 13, *c*, it is clear that (1) a much less efficient barrier diffusing region is needed to produce reductions of total intensity and (2) the intensity will remain low as long as the barrier is present.

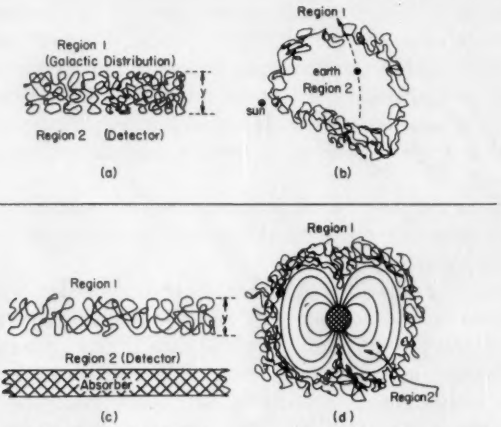


FIG. 13. Models for the diffusion of cosmic-ray particles through different configurations of irregular and twisted magnetic field regions in the interplanetary medium. These processes may produce modulation of total cosmic-ray intensity.

Now, for a diffusing barrier having the scale size of the solar system, there are no large absorbers capable of producing this effect. But if the diffusing barrier is placed around the earth overlying the earth's magnetic field, then the earth becomes the absorber (Fig. 13, *d*). This latter proposal forms the basis for the theory developed recently by Parker.[6] It is capable of explaining the sharp Forbush-type decreases, the low-rigidity cutoff and its variation with the solar cycle, the 11-year cycle of intensity, all with fields the order of 10^{-5} gauss in the barrier region. The model then becomes a local, geocentric model. This is in contrast with a barrier centered about the sun and extending throughout the solar system in all directions, i.e., a heliocentric model.

For this geocentric model the earth and its outer field are surrounded by an ionized gaseous nebula of low density retained by gravitational forces but buoyed at an equilibrium distance above the geomagnetic field by the pressures between the geomagnetic field, B(Terr.), and the fields in the ionized clouds, B(Cloud), i.e.,

$$\frac{B^2(\text{Terr.})}{8\pi} \backsimeq \frac{B^2(\text{Cloud})}{8\pi}.$$

Parker's model meets the objections raised by the Morrison model, which requires both fields 10^{-1}–10^{-3} gauss and places special conditions on the times at which particles from a solar flare may be detected at the earth.

A theory for the solar modulation of cosmic-ray intensity, based upon these principles, is consistent with and a strong argument for the indirect association found between cosmic-ray intensity variations and geomagnetic storms, since the interactions of highly ionized gas clouds of solar origin, with or without twisted magnetic fields, are expected to produce observable and transient effects on the geomagnetic field when captured by the earth. These interactions are only now beginning to be studied.

In discussing the solar-flare effect, we noted that the traversal of flare particles through the interplanetary medium must be intimately related to the theory of modulation mechanisms. To test a modulation theory, the flare particles can be used as "probes" in the interplanetary medium to determine the nature of scattering magnetic fields (or beams) which might lie between the sun and the earth. Or, by the delayed arrival of flare particles, we may place limits on the extent to which there exist diffusing clouds in the solar system beyond the orbit of the earth. Thus a test consists of a superposition of a modulation effect and a solar flare producing cosmic-ray particles. This superposition occurred under almost ideal circumstances for the flare of February 23, 1956, as shown in Figure 14.[4] Early in

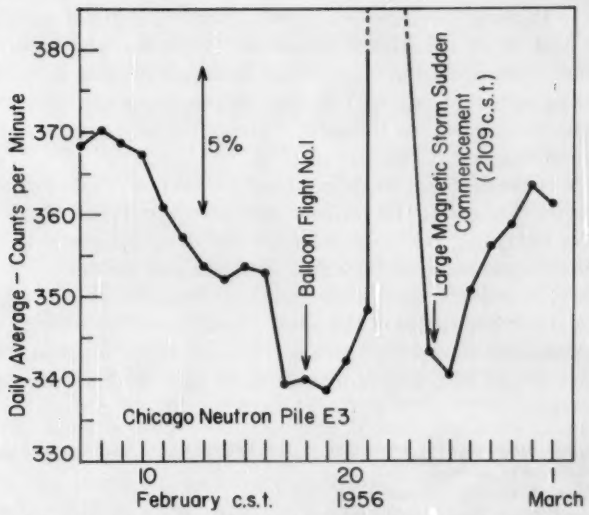


FIG. 14. A decrease of total cosmic-ray intensity similar to the event in Fig. 2 occurred in February, 1956, prior to the large solar-flare event. This superposition of modulation and production event reduces the choices for the magnetic field configurations in Fig. 13.

February the cosmic-ray intensity began to decrease, rapidly reaching a value which corresponded to the loss of 20–25 per cent of all the primary cosmic radiation reaching the earth. Then, on February 23, the sharp increase of cosmic-ray intensity arising from solar production of cosmic radiation occurred—an increase of many fold the average cosmic-ray intensity. Clearly, the particles leaving the vicinity of the flare must travel along relatively uncomplicated trajectories in order to produce the rapid increase of intensity observed at the earth. Conse-

quently, since we assume that the region within the earth's orbit was relatively free of extensive beams or clouds containing random magnetic fields, this superposition test of *modulation* and *production* favors the geocentric model.

But we also have shown that following the initial flare burst the radiation appears to disappear in such a way as to suggest extensive diffusion beyond the orbit of the earth. We may ask what happens to the outgoing ionized matter at great distances from the sun. If a galactic arm field is present, Davis has shown that the pressure of the outgoing ionized matter against the galactic arm field produces a cavity whose dimensions vary with the solar cycle.[7] However, we now believe that such a cavity will tend to be unstable with respect to penetration by the galactic field, so that only an irregular and imperfect cavity could exist. The diffusion implied by the "tail" of the solar-flare effect is the only evidence we have so far that such an irregular, cosmic-ray-diffusing region may be present beyond the orbit of the earth.

IV. *Conclusions.*—The observed nonperiodic changes of primary cosmic-ray intensity within the solar system are either the result of occasional, *de novo* production in the region of the sun or the modulation of pre-existent radiation from outside the solar system. As we have shown earlier, these variations represent changes in the primary spectrum over a wide range of particle energies involving over 95 per cent of all the cosmic radiation. Since the modulation mechanism which produces these spectrum changes has its origin in solar activity, it is clear that only during periods of extreme minimum solar activity will we be able to record briefly the galactic spectrum at the earth. Hence the implications of these studies for the origin of cosmic rays become obvious.

Irrespective of the fact that we do not have precise and detailed explanations for all the observed changes in the primary spectrum, we do feel that further work depends upon extending our knowledge of the electromagnetic properties prevailing in the interplanetary medium near the earth and the sun.

In summary, it is likely that further studies of cosmic-ray intensity variations may (*a*) extend our knowledge of the electrodynamics of the solar system and the nature of geomagnetic storms in particular, (*b*) place upper limits on the production of cosmic rays at the sun, and (*c*) enable us to gain further information on the galactic spectrum.

* Based in part upon a paper presented at the International Conference on Cosmic Rays, I.U.P.A.P., Guanajuato, Mexico, 1955.

[1] S. E. Forbush, *J. Geophys. Research*, **59**, 525, 1954.

[2] J. A. Simpson, *Phys. Rev.*, **94**, 426, 1954.

[3] J. Firor, *Phys. Rev.*, **94**, 1017, 1954.

[4] Peter Meyer, E. N. Parker, and J.A. Simpson, *Phys. Rev.*, **104**, 768, 1956; *ibid.*, **96**, 1085, 1954.

[5] P. Morrison, *Phys. Rev.*, **101**, 1397, 1956.

[6] E. N. Parker, *Phys. Rev.* **103**, 1518, 1956.

[7] L. Davis, *Phys. Rev.*, **100**, 1440, 1955.

DIRECT DETECTION OF AURORAL RADIATION WITH ROCKET EQUIPMENT

By James A. Van Allen

STATE UNIVERSITY OF IOWA, IOWA CITY, IOWA

INTRODUCTION

In 1952 a group of us at Iowa started a systematic study of the low-energy end of the primary cosmic-ray spectrum, using balloon-launched rockets for the purpose of placing the apparatus at a sufficiently high altitude to surmount the atmospheric absorption which had clouded previous work. The first series of measurements was made near Thule, Greenland, at geomagnetic latitude 88°–89°. The cosmic-ray intensity above the atmosphere there was found to be only slightly greater than we had previously found it to be at 58°.

In the course of the interpretation of these measurements we felt the need for further measurements at intermediate latitudes. Two flights of the new series made during the following summer (1953), one at 64° and one at 74°, encountered a new phenomenon above 50 km.

We have devoted two further summer expeditions to the investigation of this effect and now feel that we are beginning to have an understanding of its nature.

EXPERIMENTAL TECHNIQUE AND OBSERVATIONS

The discovery[1] of the new "soft radiation" was made with a single Geiger tube carried within the thin-walled nose shell of a small rocket. Later apparatus[2–5] has included such Geiger tubes, as well as ones shielded with additional amounts of aluminum and lead, and NaI (Tl) scintillating crystals and photomultiplier tubes. In the 1955 flights the NaI crystals were arranged beneath 0.002-inch thick titanium foil at the foreward tip of the nose shell. The apparatus is shown in Figure 1. Samples of early results are given by Meredith *et al.*[1] One of the interesting sets of results from the 1954 series is shown in Figure 2.

Most of the quantitative data obtained thus far have come from the variously shielded Geiger tubes. A summary of the maximum observed counting rates (i.e., intensities) of radiation observed at high altitudes with the unshielded tubes flown in twenty-two rockets during the summers of 1953, 1954, and 1955 is given in Figure 3. The abscissa is geomagnetic latitude. The flights have been distributed from $\lambda = 54°$ N. to 89° N. The intensity at the lowest and highest latitudes shown is believed due entirely to cosmic rays, as conventionally defined. The great peak of additional intensity in the vicinity of $\lambda = 67°$ N. is attributed to a soft radiation associated with aurorae.

In addition, we have a large amount of data from the scintillating NaI crystals which were carried on about half the flights. These data, though of a more qualitative nature, are immensely helpful in sorting out the various hypotheses which one can make concerning the nature of the radiation.

SUMMARY OF SALIENT FEATURES OF THE OBSERVATIONS

a) The latitude distribution and the temporal variability of the effect strongly suggest that it is to be associated with aurorae.

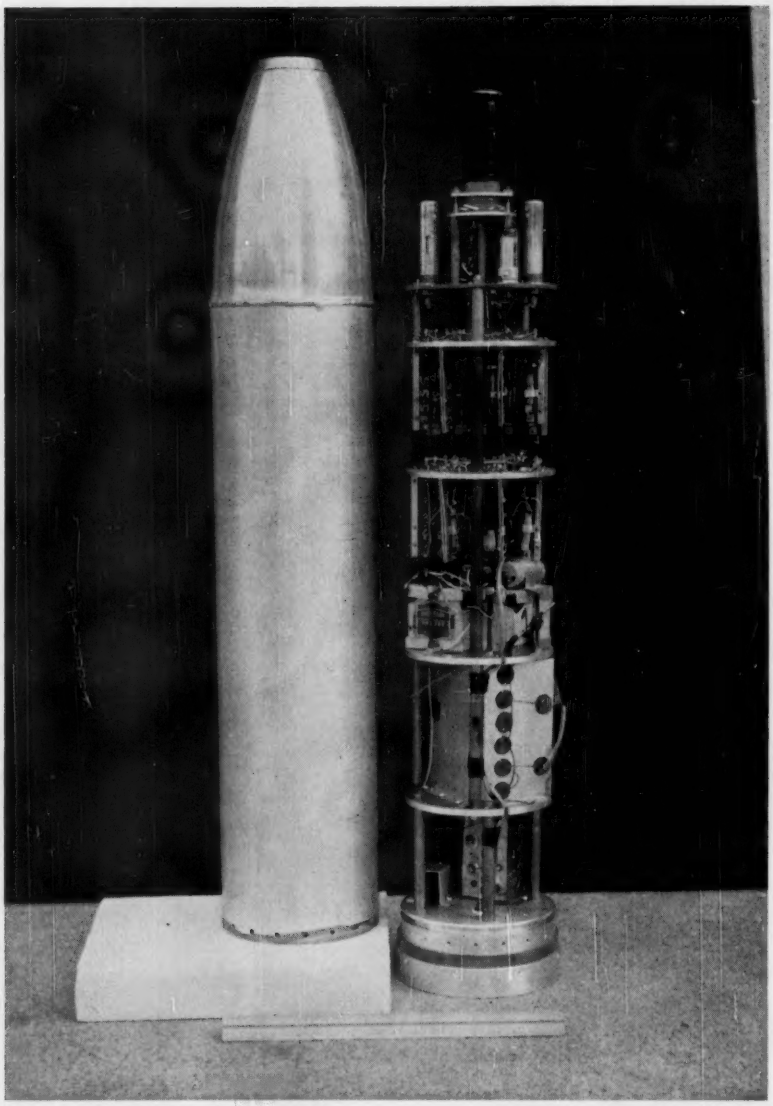


FIG. 1.—Arrangement of apparatus used in the latest series of flights (September–October, 1955). The pressure-tight aluminum nose shell on the left is placed over the apparatus on the right, and the entire assembly is attached to the forward end of a small balloon-launched rocket. The uppermost device, wrapped in black tape, is a photomultiplier tube surmounted by a NaI crystal. The shielded and unshielded Geiger tubes are the vertical cylinders at the left and right, respectively, on the next lower plate. The remainder of the apparatus consists of electronic circuits, batteries, and radio transmitter.

b) The radiation is quite soft (by cosmic-ray standards), being completely or nearly completely absorbed (when found to be present at all) by amounts of material ranging from several gm/cm² to several hundred mg/cm² of air and/or aluminum and being attenuated by a factor ranging from 3 to greater than 50 by 150 mg/cm² of lead.

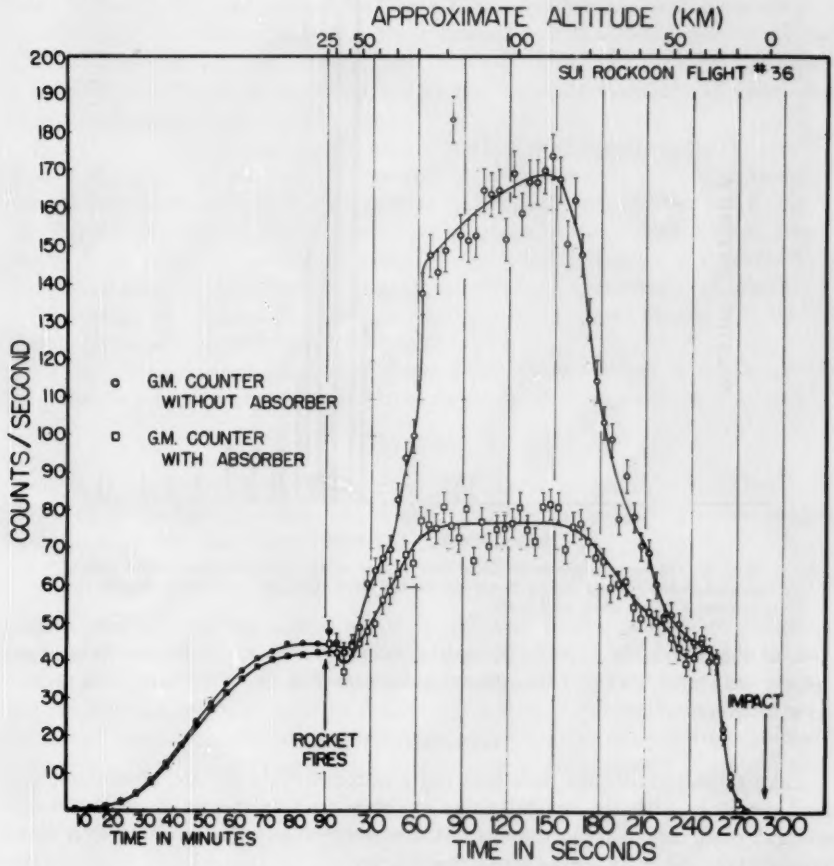


FIG. 2.—A typical set of raw flight data observed with two Geiger tubes, one without absorber and one incased in an absorber of 150 mg/cm² of aluminum. The normal cosmic-ray counting rate is about 39 counts/sec at altitudes above 50 km. The additional counts are due to the soft radiation.

c) Referring to the crystal measurements which give absolute energies dissipated in the crystal, we have observed no case in which resolved pulses corresponding to greater than 200 kev occur (except for the expected number of cosmic-ray pulses), *even though* there is a simultaneous occurrence of a very large counting rate in the more heavily shielded Geiger tubes. We do observe in the output of the crystal a marked electronic pile-up of small pulses—an effect referred to in more detail below.

d) The "wings" of the counting-rate versus time curves are in all cases "regular" in character and are believed attributable to atmospheric absorption. The detailed features of the middle portions of the curves are believed due to a combination of three possible causes: (i) spatial inhomogeneities, (ii) fluctuations of intensity with time, and (iii) changing angular aspect of the equipment with respect to anisotropic radiation.

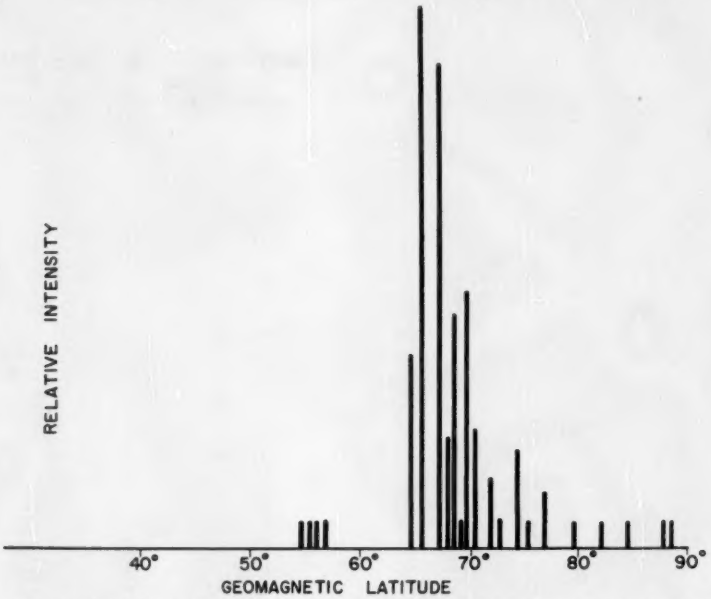


FIG. 3.—The geomagnetic latitude dependence of the maximum counting rate of the unshielded Geiger tubes flown on twenty-two different occasions during the summers of 1953, 1954, and 1955.

INTERPRETATION

Taking the Geiger tube data and the absorption data for air, aluminum, and lead, together with the crystal data, we have an assemblage of crude "range-energy" observations. On the basis of this assemblage we have sifted over three possible types of radiation as broad possibilities.

a) *Protons.*—By the absorption data it is seen that protons must have energies of the order of 20 Mev. But no pulses as large as one-hundredth of this size were observed in the crystals. Hence we believe that the observed effects certainly cannot be due to protons (at least not directly).

b) *Electrons.*—Electrons must have about 2 Mev. energy to have the typically observed penetrability. But, again, no resolved pulses in the crystals larger than one-tenth this size were observed. Hence we believe that the directly observed effects did not consist of counting primary auroral electrons.

c) *X-rays.*—It appears that X-rays whose energies at various times lie in the range 10–100 kev are consistent with all observed data.

LABORATORY SIMULATION

Actual flight apparatus has been set up in the laboratory and exposed to X-ray beams of arbitrarily adjustable characteristics. A sample run goes as follows: We adjust the accelerating voltage on the X-ray tube until the ratio of the counting rates in the shielded and unshielded Geiger tubes matches the value observed in a chosen flight. We then raise the beam intensity in the X-ray tube until the absolute flight rates are reproduced. All this is, of course, readily done and yields a rough determination of the X-ray energy and intensity necessary to match the flight observations with the Geiger tubes.

In addition, and of special significance, is the fact that in performing the test described, the output of the crystal *automatically* has a close quantitative resemblance (i.e., of electronic pile-up of small pulses) to the performance observed in flight. Further details, such as the very sensitive response of the crystal pile-up output to the X-ray intensity, are reproduced, as is the great variability of response to a directional beam of X-rays as the angular aspect of the apparatus is changed.

This program of laboratory simulation has not yet been carried out for all flights, but no inconsistencies have yet been found.

Hence it appears that this line of study gives strong support to the X-ray interpretation which has been made purely by study of the flight data.

INTENSITY

On the basis of Geiger tube efficiencies for X-rays, we estimate that the X-ray intensity is of order of magnitude 10^3–10^5 photons/cm^2 sec, the photons having energies of 10–100 kev.

PRESUMPTION

Since the presence of the observed X-radiation in the atmosphere shows a latitude distribution strongly resembling that of aurorae as observed by other means, it is presumed that the X-rays are associated with primary auroral radiations but are not themselves primary in nature. It is then of interest to see the reasonability of presuming the observed X-rays to be bremsstrahlung from primary electrons. This line of thought leads to estimates of the intensity of auroral electrons as

$$10^6\text{–}10^8 \text{ electrons/cm}^2 \text{ sec},$$

the electrons having kinetic energies of the order of 10–100 kev. The corresponding energy flux is 10^{10}–10^{12} ev/cm^2 sec, or

$$0.01\text{–}1 \text{ erg/cm}^2 \text{ sec}.$$

Auxiliary checks on this presumption are the following:

a) The estimated amount of incident electron energy (much of which goes into optical excitation of the atmospheric gases) appears to be in plausible agreement with other estimates of the optical energy of visible aurorae ($\sim$1 erg/cm^2 sec for bright aurorae).

b) Electrons of 10–100-kev kinetic energy would penetrate to 90–110-km. altitude in the atmosphere, a result again in agreement with optical observations.

It is also of interest to examine the possibility that the X-radiation may be bremsstrahlung from *protons*. In this case an incident energy of protons of the order of 1 joule/cm^2 sec is required. Such a magnitude of energy flux (considerably greater than that of sunlight) is regarded as totally unreasonable on the basis of its enormous consequences on the entire upper atmosphere (above 100 km.) in which it would be absorbed.

NEED FOR FURTHER WORK

a) Thus far, all the flights have been made during daylight hours and/or through overcast skies. Aurorae have been visually observed during some nights of the general period of the flight operations. But we have obtained no directly correlated observations (either visually or by radio methods) during flights which either have or have not revealed the presence of the soft radiation.

It is planned to make such co-ordinated observations at Fort Churchill, Manitoba during the International Geophysical Year period.

b) The direct verification of the primary electron hypothesis developed above is well within the range of experimental feasibility using a simple windowless magnetic spectrometer carried in a small rocket. This has not yet been done but is planned for the next series of flights. (The magnetic rigidity of the hypothesized electrons is $H\rho$ = 500–1,000 gauss cm., a suitable value for a simple spectrometer.)

REMARKS

a) The magnetic rigidity of the hypothesized electrons is distinctly below the Störmer geomagnetic cutoff for independently traveling electrons from a remote source, and it seems much more likely that they arrive at the top of the atmosphere in a manner resembling that of Chapman-Ferraro-Martyn.

b) Emphasis has thus far been given, in our work, to the direct establishment of the physical nature of the radiation, but it is now becoming acutely clear that the interrelationship of this type of measurement to other types of visual, solar, and magnetic measurements must be established.

c) The X-rays encountered at relatively low altitudes (40–70 km.) are thought of as the bremsstrahlung from electrons which have been stopped at 90 km. or above in the atmosphere. But when the apparatus is at, say, 100 km., it is to be supposed that the primary auroral electrons are incident on the walls of the apparatus themselves and are generating bremsstrahlung locally.

The assistance of a joint program of the Office of Naval Research and the Atomic Energy Commission is gratefully acknowledged.

[1] L. H. Meredith, M. B. Gottlieb, and J. A. Van Allen, "Direct Detection of Soft Radiation above 50 Kilometers in the Auroral Zone," *Phys. Rev.*, **97**, 201–205, 1955.

[2] F. B. McDonald, R. A. Ellis, and M. B. Gottlieb, "Rocket Observations on Soft Radiation at Northern Latitudes," *Phys. Rev.*, **99**, 609, 1955 (abstr.).

[3] J. A. Ellis, M. B. Gottlieb, and L. H. Meredith, "Double Geiger-Tube Study of Soft Radiation at Northern Latitudes," *Phys. Rev.*, **99**, 609, 1955 (abstr.).

[4] R. A. Ellis, "High-Altitude Rocket Measurements with Scintillation Counters at Northern Latitudes," *Phys. Rev.*, **99**, 609, 1955 (abstr.).

[5] J. A. Van Allen, "Interpretation of Soft Radiation Observed at High Altitudes in Northern Latitudes," *Phys. Rev.*, **99**, 609, 1955 (abstr.).

PROBLEMS OF AURORAL MORPHOLOGY

By C. T. Elvey

GEOPHYSICAL INSTITUTE, UNIVERSITY OF ALASKA, COLLEGE, ALASKA

In the past the problems of auroral morphology were involved with the descriptions of individual auroral forms—their heights, dimensions, motions, etc. An example of such arcs is shown in Figure 1. Now the problems that confront us concern the broad scale of the auroral display rather than the individual components. Our knowledge concerning the broader aspect of auroras depends upon meager and scattered evidence. A large part of the information is statistical in character, giving, for example, the average distribution of auroral displays over the earth and the average frequencies of occurrence with respect to time of night, to the seasons, or to the phase of the sunspot cycle. Facts concerning the nature of a given auroral display and how it changes with time are lacking. In addition, we need to know the physical processes associated with the auroral display. These physical processes reveal themselves in many different terrestrial phenomena, such as magnetic storms, ionospheric disturbances, earth current disturbances, ionospheric absorption, and radio noise from auroras. Each of these phenomena will serve as a tool for untangling the physical processes involved. Statistical relationships among the phenomena are known, but detailed studies of their relationships have only been started. These problems constitute a large part of the auroral program of the International Geophysical Year, 1957–1958.

To serve as background, we shall review a general interpretation of auroral and related phenomena. This interpretation may not be acceptable to everyone in all its details, but there is a considerable amount of observational evidence to substantiate it. It may be thought of as a working hypothesis that has been used in designing the United States auroral program for the International Geophysical Year.

A neutral stream of charged particles is emitted, or propelled, from an active area on the sun. The composition of the stream should be essentially that of the sun's atmosphere, that is, mainly protons and electrons, with a small percentage of heavier atoms. Evidence indicates that the stream or cloud of gas is propelled with sufficient velocity to reach the earth in approximately 24 hours. As the stream or cloud sweeps across the earth, the magnetic field of the earth deflects and captures some of the charged particles. The particles bombard the atmosphere in regions surrounding the geomagnetic poles, the auroral zones. The particles spiraling down the magnetic lines of force penetrate to depths in the atmosphere corresponding to their energy. The average base of most auroral forms is 100 km. above the surface, and it is believed that a proton of 400 kev. energy will penetrate to that depth in the atmosphere. Both direct and indirect effects will result from this bombardment of the atmosphere.

The most important direct effect of this bombardment of the atmosphere is the polar aurora, wherein the particles excite the outer shells of electrons of the atmospheric gases to radiate light. Second, the particles ionize the gas, leaving trails of electrons that are detected by radar and radio waves and have become known as "radar auroras." A third direct effect, and one which promises to become an extremely important one, is the discovery of X-rays in auroral regions by the rocket group of Iowa State University working under Van Allen.[1] The X-rays have

been detected by using rockoons (balloon-launched rockets) in a survey across the auroral zone from the North Atlantic Ocean into the Davis Straiht, west of Greenland. Other possible direct effects of the bombardment are the heating of the atmosphere in the auroral zone, causing disturbances in the atmospheric circulation, and the emission of radio noise by the charged particles moving in the earth's magnetic field.

FIG. 1.—System of auroral arcs.

A very important and practical secondary effect, the nondeviative absorption of radio waves in the ionosphere, has been discussed by Chapman and Little.[2] The X-rays excited by the incoming charged particles will ionize the atmospheric gases, mostly at the levels below the aurora and in layers, owing to the exponential increase of density. The authors show that the absorption will be present both by day and by night, even though there is a minimum of auroral activity near the middle of the day, as shown by radar techniques. Although the X-rays will be far lower in intensity during the day, the full effect is felt because the sunlight prevents the attachment of electrons and oxygen atoms. At night the attachment process can proceed uninhibitedly. Consequently a far greater supply of X-rays is required to produce the same amount of absorption. It also seems reasonable that the sporadic E clouds associated with auroras may be explained by the X-ray phenomenon.

Other secondary, or perhaps we should call them tertiary, effects result from the moving of the ionized gases by the tides, winds, and turbulence of the upper atmosphere. These effects will evidence themselves as ionospheric storms, increased scintillation of radio stars, and other ionospheric phenomena associated with auroral displays. The moving charges will produce electric currents in the upper atmosphere that will reflect as induced currents in the earth, the earth potential disturbances. Likewise, the magnetic fields associated with the atmospheric electric currents will produce the disturbances associated with magnetic storms.

Returning now to the description of an auroral display and its changes or development with time, we find that most descriptions are from the local, or small-scale, point of view. It is possible to obtain a fairly good description of the development of a small auroral display from observations at one point; however, attempting to generalize with regard to the whole display or to extensive displays is very hazardous. Under the best of observing conditions, the observer can see auroras over an area whose radius is approximately 1,000 km. Nevertheless, the resolution of auroral forms at distances greater than 500 km. is rather poor; hence it is safer to restrict interpretations to those within a radius of 500 km. The above is based on the assumption that the height of the base of the aurora is 100 km.

Recently, Heppner[3] made a study of the development of auroral displays along the magnetic meridian through College, Alaska. He assumed the height of the lower edge of the auroral forms and was thus able to determine from measurements of the elevation above the horizon the geomagnetic latitude of the place where the auroral form was incident. He illustrated the development of the auroral display graphically by plotting the auroral forms in coded symbols on a chart. The ordinate is geomagnetic latitude, and the abscissa is time. Figure 2 is the diagram for the aurora on the night of January 25–26, 1952.

We have expanded this idea to show the distribution of auroras in two dimensions. Visual observations from five stations are combined to give synoptic maps of the distribution of auroras over Alaska for each quarter-hour. The synoptic maps for the beginning of each hour during the night of March 26/27, 1954, are shown in Figure 3.

Observers working in the region of the auroral zone, where auroras frequently not only cover the entire sky but also are often very active, find it difficult to record visual observations satisfactorily. Therefore, we have investigated the possibility of making the observations photographically. An idea of Gartlein[4] for photographing the auroras over the entire sky has been adopted. It is termed the "all-sky" camera and consists of a convex mirror with a 16-mm. camera mounted above it. The photographs are images of the entire sky as seen in the convex mirror. Exposures are made at 1-minute intervals and thus give comprehensive records of auroral displays. A typical photograph of an arc system is shown in Figure 4. This photograph was made with the earlier model of the camera, in which the duration of the exposure was 54 seconds. The remaining 6 seconds of each minute was used to transport the film for the next exposure. The results of the measurements of Figure 4 are shown graphically in Figure 5. It is to be noted that when arcs are photographed, it is frequently possible to trace them to the visible horizon. Knowing the elevation of the visible horizon from actual measurement, it is very

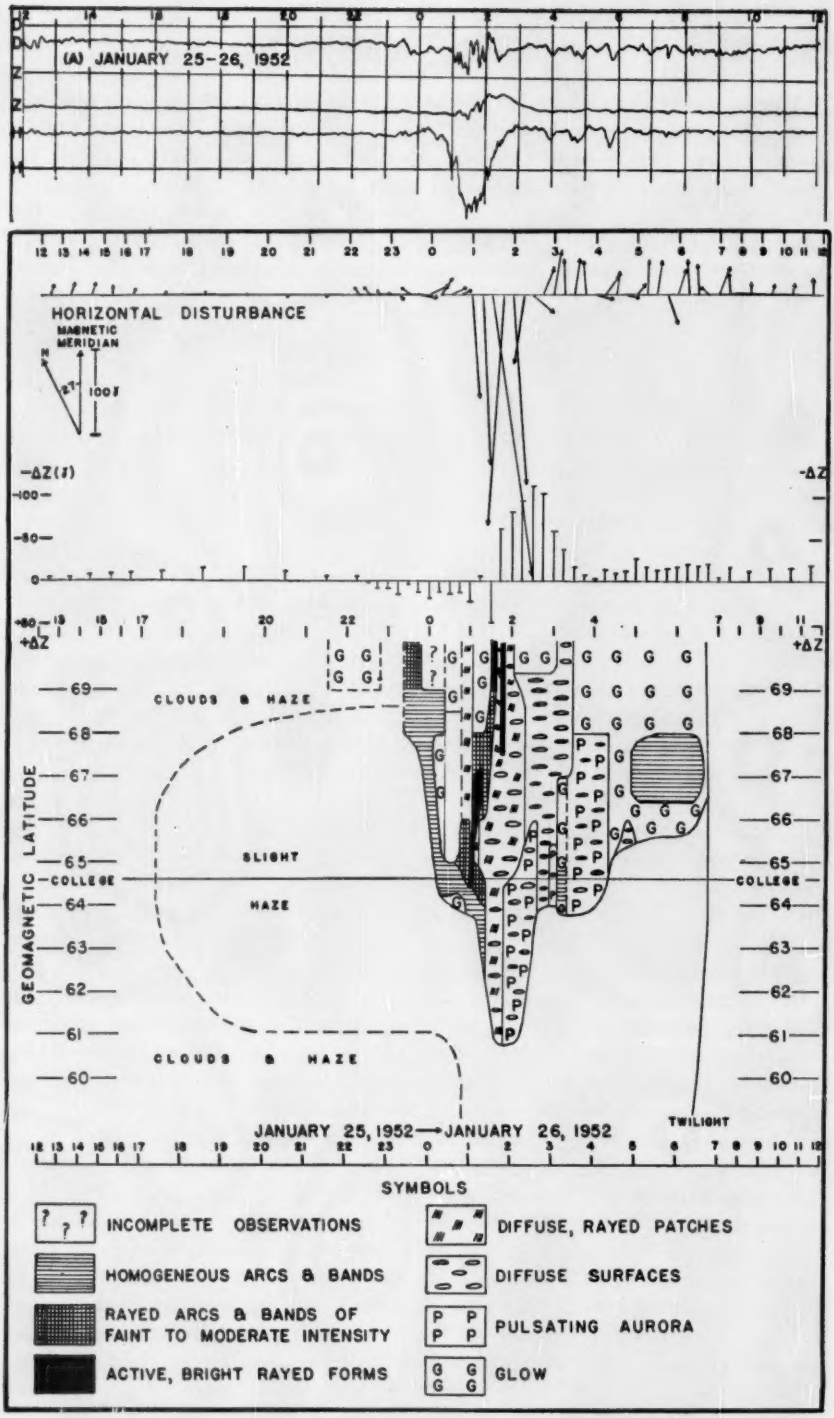


FIG. 2.—Auroral display of January 25–26, 1952 (after Heppner).

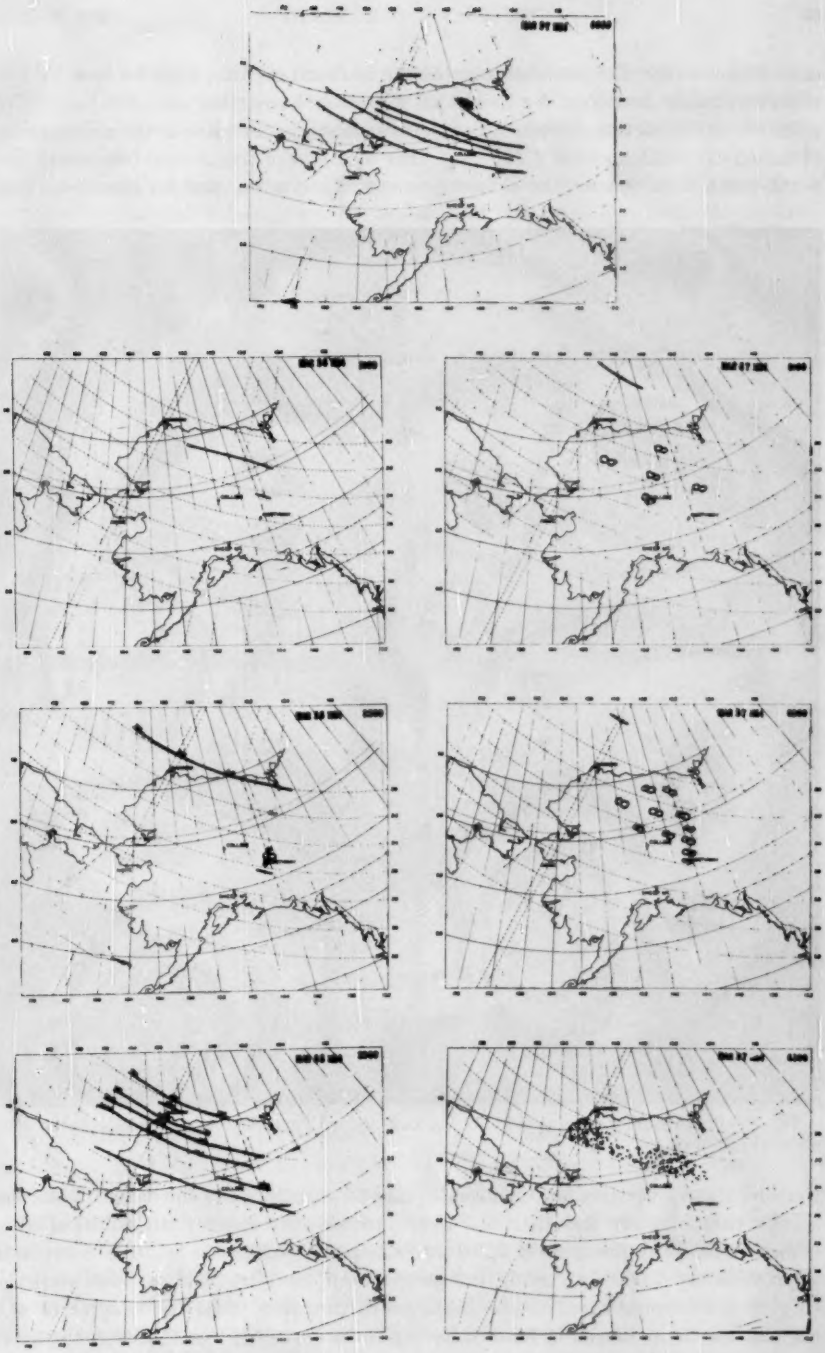


FIG. 3.—Distribution of aurora over Alaska, March 26–27, 1954.

easy to determine the farthest point of the arc's extent that may be seen by the observer, again assuming the height of the lower border at, say, 100 km. The azimuth can be measured quite accurately, so that the positions of the arcs may be obtained up to distances of 1,100 km. The locations of the auroral features in the north-south direction require a measurement of elevation, and for elevations less

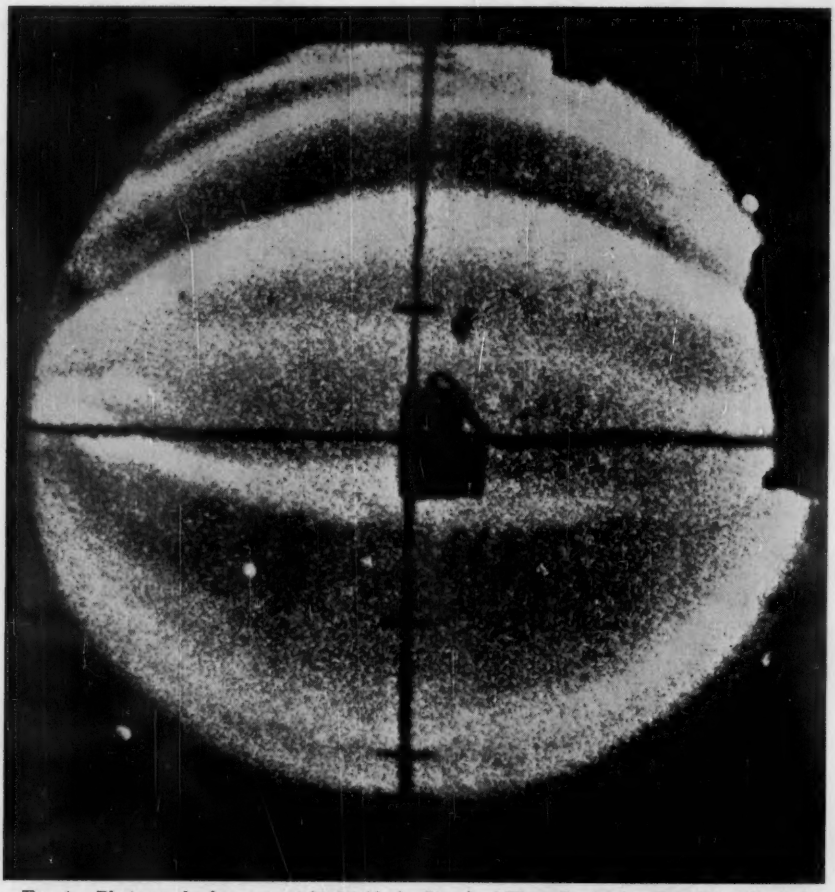


FIG. 4.—Photograph of auroras at 21:34 Alaska Standard Time, September 30, 1954, taken with an all-sky camera.

than 10 degrees the accuracy is not adequate to warrant using the data. Thus the all-sky camera gives the distribution of auroras over roughly an elliptical area whose extent is approximately 2,200 km. along a geomagnetic latitude circle and approximately 1,000 km. along the geomagnetic meridian. If cameras are used to give overlapping areas in the north-south direction, then the heights of the auroras can be measured instead of having to be assumed. In that way the positions of the auroras will be more accurately located.

The all-sky cameras are also useful for the determination of the times of breakups of arcs. In Figure 6 a pseudo-breakup is shown. In a system of arcs the pseudo-breakup is defined as the disruption of one of the arcs near to the center of the auroral zone.

We believe that during the last four or five years in Alaska we have been able to derive a fairly reliable description of the southern portion of small auroral displays, that is, ones associated with magnetic bays. However, the longitudinal extents of the displays are greater than can be seen from our several observing stations. The display begins to form in the northern part of Alaska or to the north of Alaska in the evening hours, say 20:00 to 21:00 hours local time. The first evidence of the

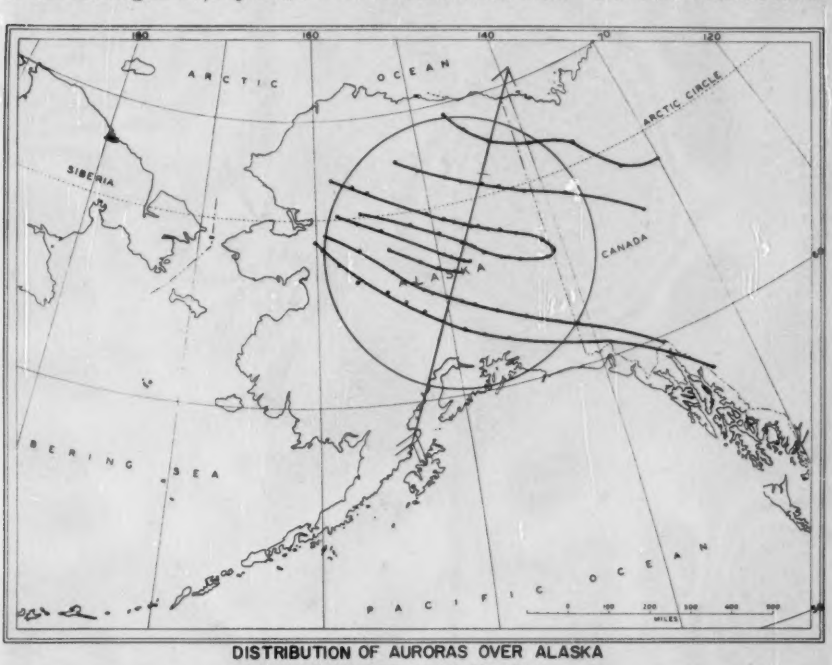


FIG. 5.—Distribution of auroras over Alaska on September 30, 1954, at 21:34 Alaska Standard Time.

display is a faint homogeneous arc. In the simplest case a display may consist of only one arc; however, most displays consist of a number of arcs. Sometimes the arcs are so close together that the group is called a "multiple arc"—two to six arcs within a distance of about 25 km. along the geomagnetic meridian. The arcs are quite accurately parallel to each other, extending from horizon to horizon, a distance of over 2,000 km. The arcs are parallel, or nearly parallel, to a geomagnetic latitude circle. Another striking fact is the thinness of some of the arcs. On one occasion the width of an arc passing through the magnetic zenith (at this position only the bottom edge of the arc is seen) was estimated to be one-fourth the diameter of the moon, which was nearby. Hence its thickness was not more than 250 meters.

The above facts certainly represent precision focusing of the incoming charged particles.

The arc or arc system gradually moves southward, reaching the zenith at College around midnight. During the years 1953–1943 this seemed to be about the southern limit for the arcs. They would halt at or near the zenith, leisurely moving to and fro in unison. Throughout its lifetime the arc varies in intensity, usually very slowly.

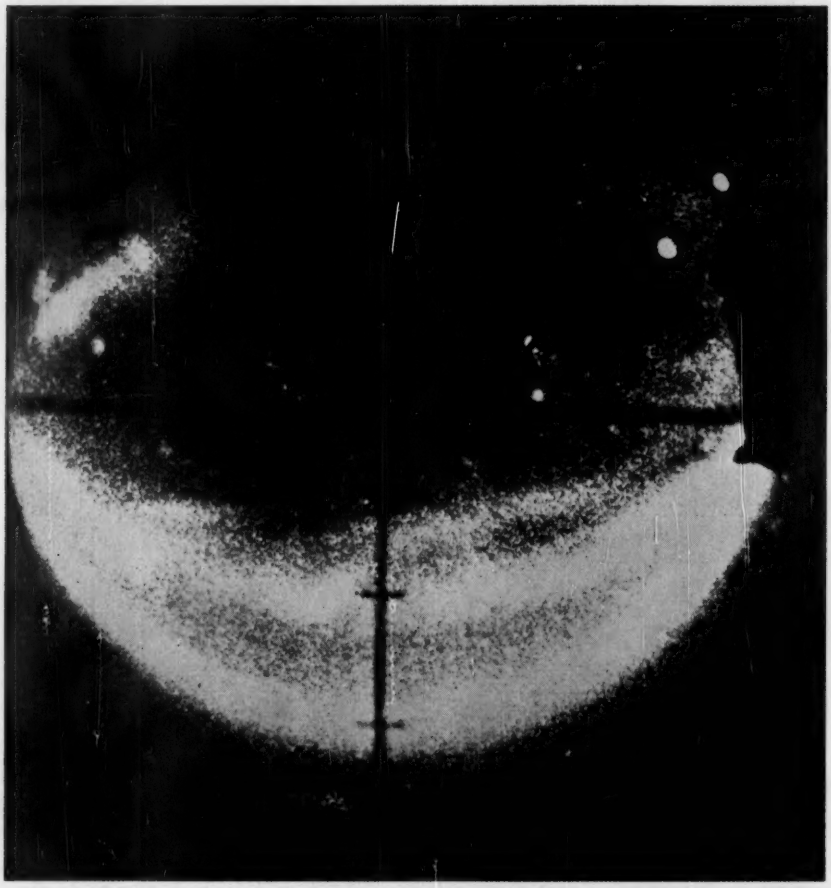


FIG. 6a.—Photograph of auroral display showing partial breakup of arc system on October 1, 1954, 01:25 Alaska Standard Time.

If there are several arcs in the system, one or more of the ones closer to the auroral zone may break up, with the pseudo-breakup sometimes forming a rayed arc or breaking into bands and various rayed structures. These pseudo-breakups appear to coincide with the minor bays that give the irregularity to the main bay in the horizontal component of the magnetic record.

The final breakup of the display appears to be at the time that the arc farthest from the auroral zone disrupts. This time is a very critical phase in the auroral display. The arc brightens and quivers just before disrupting into rays, bands, and draperies that spread rapidly over a large part of the sky as though the incoming particles has been defocused. Very shortly thereafter the display may change to flaming aurora and pulsating surfaces. The flaming aurora seems to be a transitory

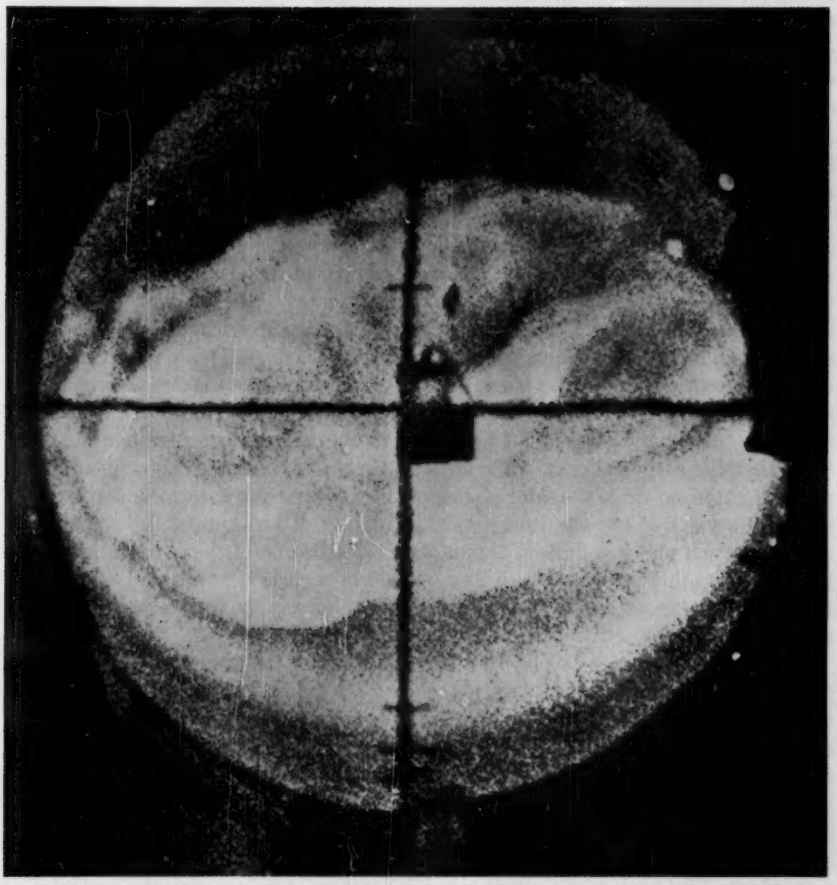


FIG. 6b.—Photograph of auroral display showing partial breakup of arc system on October 1, 1954, 01:26 Alaska Standard Time.

stage. The pulsating surfaces usually last for a considerable time and gradually change into diffuse surfaces. The diffuse surfaces may recede toward the auroral zone, becoming more diffuse and fading, thus terminating the display.

The descriptions of auroral displays by observers inside the auroral zones indicate that the display develops in the same manner with respect to the center of the auroral zone; however, since we do not have simultaneous observations from both

sides of the auroral zone, we do not know whether the auroral display sometimes develops to the south of the zone and at other times to the north or whether it spreads in both directions from the center of the zone. The Canadian observers[5] have shown that the center of the auroral zone shifts with respect to the season and with respect to the sunspot cycle or perhaps the sunspot activity.

Thus one of the basic problems in auroral morphology is to determine the extent of a single small display, both in latitude and longitude, as a function of time after the start of the display. Since the smaller displays seem to cluster around local midnight, does the display move westward in phase with the sun? Or does the display essentially stand still in longitude and occur at different local times around the auroral zone? We know statistically that auroras occur in the southern auroral zone in the same manner as in the northern, but we do not know whether there is a southern counterpart of the small auroral display of the northern zone. And if such a counterpart exists, does it stay in phase throughout the development?

The matter of the hydrogen spectrum in the aurora offers another interesting problem. Meinel[6] has shown that the Doppler shift of the hydrogen lines corresponds to velocities up to 3,000 km/sec when viewed at the magnetic zenith, that is, the direction of travel down the magnetic line of force. When viewed perpendicular to the magnetic lines of force, the hydrogen lines are broadened but do not have a Doppler shift of the line as a whole. Meinel has found that the hydrogen spectrum is present during the arc phase of the auroral display but is not present in the rayed structures. Meinel and Fan[7] conclude from laboratory studies of the excitation of atmospheric gases in a low-pressure discharge tube using protons and electrons that the earlier phases of the auroral display have spectra corresponding to proton excitation, while the later phases of the display have spectra similar to electron excitation. It appears that the critical phase of breakup of the arc system may indicate a major change in the method of excitation of the auroral spectrum. However, it is difficult to see why there should be a close relationship between the stoppage of the protons and the change from the homogeneous and essentially quiet auroral forms to the active phase of the display.

The part that the protons play in the development of the auroral display is therefore an important problem. Are the protons the primary exciting agency, or do they merely serve as a trigger mechanism or catalyst? Much more evidence is required concerning the distribution of hydrogen during the auroral display. Is it possible that the protons are concentrated in the arcs during the early, or quiet, phase of the display? Are they precisely focused, and does the breakup phase of the display represent a defocusing of the proton beam? This suggestion could be checked by measuring the intensity of the hydrogen spectrum integrated over a rather large solid angle of the zenith sky as a function of time during the auroral display. A recording photoelectric photometer combined with a filter combination of sufficient spectral purity or a monochromator would be an excellent instrument for this purpose. Another important question might be answered at the same time, namely, are the intensity variations of the hydrogen spectrum closely correlated with the variations in intensity of the auroral light? This situation might be expected if the excitation of the aurora resulted from proton bombardment of the atmosphere.

Other problems that should be taken into consideration in our discussions of auroral morphology are the abnormal cases. This need was brought to our attention by the appearance of the red auroral display that occurred about 55 hours following the solar flare of February 23, 1956. The red display has been described by Leinbach[8] and also by Elvey,[9] who was able to determine its position in the atmosphere from many observers in Alaska, Western Canada, and the northwestern United States. The red arc is a rare phenomenon; Størmer[10] records reports of only four since 1911. The red arc of February 25, 1956, extended from the Bering Sea near the Siberian coast across the North Pacific Ocean and the states of Washington and Montana. East of Montana it was cloudy, but Gartlein reported an enhancement of the red auroral lines, using a patrol spectrograph at Ithaca, New York, which indicates that the arc probably extended that far. The arc was essentially along geomagnetic latitude 54° N. The height of the lower edge of the arc was found to be at least 335 km. Only one of the arcs reported by Størmer was measured; the height was found to be 250 km.

The red arc seemed to fade rather than break up, and for many minutes following its disappearance there were scattered red auroral forms over much of the sky as seen from College, Alaska. Most of the forms were red diffuse surfaces; however, one sharply defined red ray was seen. The evidence indicates that the red display passed through phases somewhat similar to a normal auroral display of low intensity.

Since the red arc had a rather sharply defined lower border at a height of 335 km. or greater, the indication is that the charged particles penetrating the atmosphere must have possessed very little kinetic energy in comparison with the charged particles that are involved in the usual auroras. Sugiura[11] made calculations of the penetration of protons in the atmosphere, using the standard atmosphere from rocket data. He found, for example, that a 400-kev. proton can penetrate to a height of 100 km., whereas a proton with only 1 kev. can penetrate to about 140 km. If we are to assume that the excitation of the red auroral arc is a result of the bombardment of the atmosphere by protons, then the protons must have had extremely low kinetic energy.

That is a problem for the theorists. We shall offer two suggestions, however. The first is the reconsideration of the ultraviolet-light theory of aurora proposed by Maris and Hulbert.[12] An intense solar flare such as the one that occurred slightly more than two days before the red auroral arc could give rise to a spray of gases from the upper atmosphere, centered at the subsolar point. The gases would fall back under gravitational forces, but the ionized particles would fall back to earth following the magnetic lines of force and re-enter the atmosphere in polar regions. The second suggestion is to consider the influence of a marked increase in ionization of the outer regions of the atmosphere. The resistance of an ionized gas to the passage of protons should be orders of magnitude greater than that of a neutral gas. Under normal conditions the upper atmosphere is predominantly not ionized. The first suggestion would appear to be attractive as far as the red aurora of February 25, 1956, is concerned. There are, however, other bits of evidence that point to the increased ionization of the upper regions of the atmosphere as a factor in slowing the bombarding protons.

Størmer[13] has described another rare form of auroral arc. Faint isolated arcs have been observed to occur following an intense display lasting two or three days. The measured heights of the bases of these arcs are around 200 km. In this case also, the greater heights are associated with periods when the upper atmosphere must have had increased ionization because of the auroral activity.

Normal auroras are observed to be in the range between 90 and 120 km., with the average height of the lower border being 100 km. (The heights of the various types of arcs are shown schematically in Fig. 7.) There is a group of auroras, how-

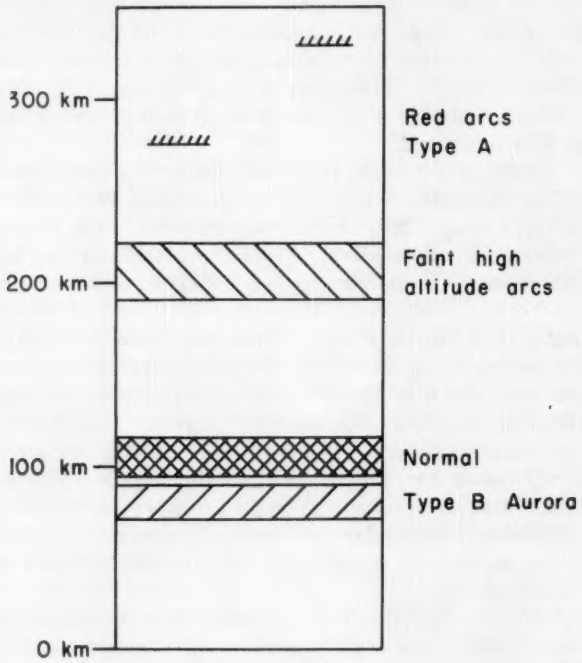


FIG. 7.—Heights of auroral arcs.

ever, whose lower borders have been measured in the range from 70 to 90 km. These are the Type B red auroras, which are characterized by red or purplish-red lower borders. The greatest frequency of appearance of the Type B red auroras is during sunspot minimum, a time when the upper atmosphere is at a low degree of ionization. Another characteristic of these auroras is that the light is much more concentrated in the lower portion of the auroral forms, with little or no extension to high altitudes.

The Type A red auroras have long extensions of the arcs and rays into the upper atmosphere, being red in the upper portions. The Type A red auroras appear during sunspot maxima, when the ionization of the upper atmosphere is greatly increased. The appearance of the auroral forms indicates that the exciting particles

are being slowed down much higher in the atmosphere than occurs at sunspot minima, when the Type B red auroras are predominant.

From the general physical interpretation of the observations of unusual heights of aurora, it seems probable that a consideration of the degree of ionization of the upper atmosphere may account for the wide range found in auroral heights. For this reason the degree of ionization may become one of the problems of auroral morphology.

The auroral program for the International Geophysical Year, 1957–1958, was designed with many of the above-mentioned problems in auroral morphology in mind. Hence we may expect to have answers for some of the problems in the next two or three years.

[1] J. A. Van Allen, *Phys. Rev.*, **99**, 609, 1955.
[2] S. Chapman and C. G. Little, *J. Atm. and Terrest. Phys.* (in press).
[3] J. P. Heppner, *J. Geophys. Research*, **59**, 329, 1954.
[4] C. W. Gartlein, *Nat. Geographic Mag.*, **92**, 683, 1947.
[5] F. T. Davies, *Transactions of the Oslo Meeting, August 19–28, 1948* (Washington: International Union of Geodesy and Geophysics, 1950).
[6] A. B. Meinel, *Astrophys. J.*, **113**, 50, 1951; **114**, 431, 1951.
[7] A. B. Meinel and C. Y. Fan, *Astrophys. J.*, **115**, 330, 1952.
[8] H. Leinbach, *Sky and Telescope*, **15**, 329, 1956.
[9] C. T. Elvey, Seventh Alaska Science Conference, Alaska Division, AAAS, Juneau, September 26–29, 1956.
[10] C. Størmer, *The Polar Aurora* (Oxford: Clarendon Press, 1955).
[11] M. Sugiura, *Proceedings of the Third Alaska Science Conference, September 22–27, 1952*, p. 32.
[12] H. B. Maris and E. O. Hulburt, *Phys. Rev.*, **33**, 412, 1929.
[13] C. Størmer, *Photographic Atlas of Auroral Forms and Scheme for Visual Observations of Aurorae* (Oslo: International Geodetic and Geophysical Union, 1951).

SOLAR INFLUENCES ON GEOMAGNETISM

BY JULIUS BARTELS

GEOPHYSIKALISCHES INSTITUT, UNIVERSITÄT GÖTTINGEN, AND MAX-PLANCK-INSTITUT FÜR PHYSIK DER STRATOSPHÄRE

(*Read by S. E. Forbush*)

Geomagnetic records provide information respecting various solar influences upon the earth: Corpuscular radiation is effective in producing immediate geomagnetic activity, as well as in the aftereffects known as "post-perturbation." Wave radiation is effective in causing diurnal variations of the quiet-day types, varying slowly with solar activity in the 11-years cycle, or in the solar rotation period, and occasionally varying rapidly, during solar flares.

Magnetograms are of particular value for the continuous measurement of that part of solar *corpuscular radiation* (P) which has travel times between about 20 hours (in the production of sudden commencements of storms) and about 3 or 4 days (in the medium geomagnetic activity exhibiting 27-day-recurrence tendencies). The 3-hour-range K-index has been devised to measure the effects of P locally, including its considerable systematic changes with local day-time. A standardized abstract of the K-indices of many magnetic observatories, called the "planetary

three-hour-range index" Kp, provides a summary for the earth as a whole. (See Figure 1.) K-indices and Kp are available for the Second International Polar Year, 1932–1933, and currently since the beginning of 1937. Indices for complete days, such as the equivalent amplitude Ap or the planetary character figure Cp, are derived from Kp. Monthly tables and 27-day-recurrence diagrams for Kp are published around the twentieth of each month, and international data on magnetic disturbances appear quarterly in the *Journal of Geophysical Research*.

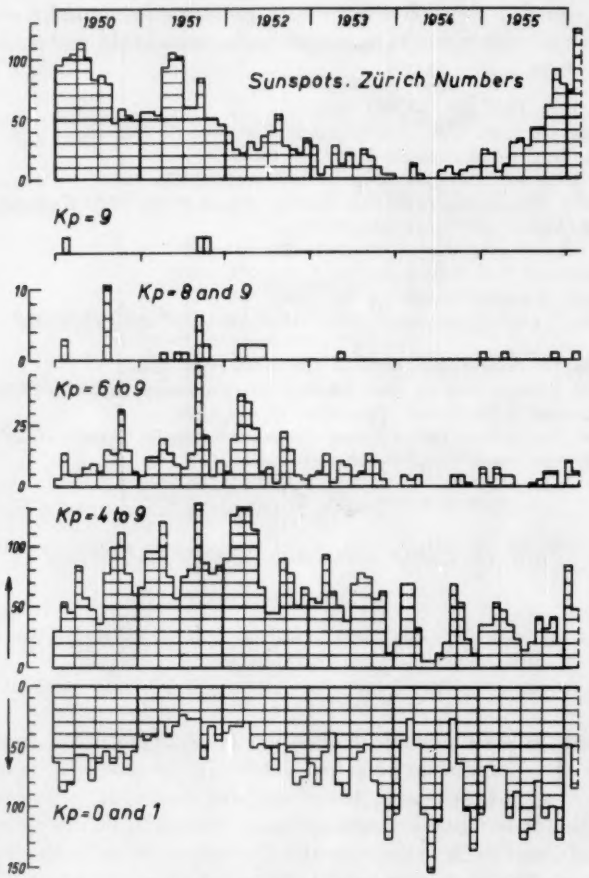


FIG. 1.—Monthly numbers, January, 1950, to February, 1956, for Zürich relative sunspot numbers and frequencies of magnetic planetary 3-hour-range indices, Kp.

These K-indices, apart from their immediate value in geomagnetic work itself, are being used as a reliable measure of solar corpuscular radiation in studies on aurorae and airglow, in work on the ionosphere, on meteors, on comets, on cosmic radiation, etc. As to the inference to be drawn about solar events themselves, it is clear

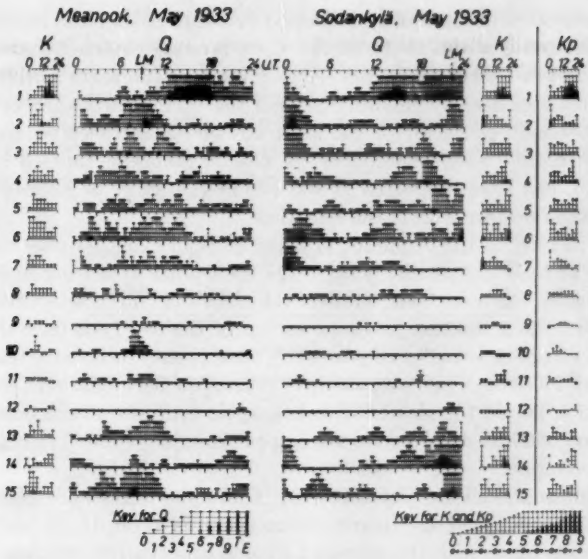

Meanook, May 1933
Sodankylä, May 1933
K
Q
Kp
LM
ut
Key for Q
Key for K and Kp

Fig. 2*a*

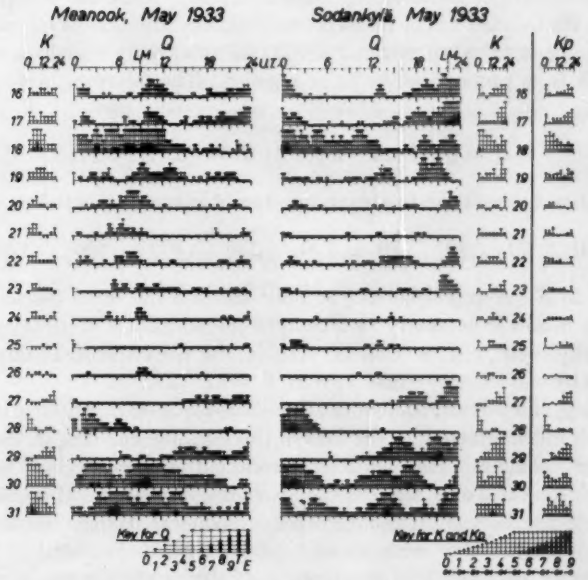

Meanook, May 1933
Sodankylä, May 1933
K
Q
Kp
LM
ut
Key for Q
Key for K and Kp

Fig. 2*b*

that the significance of Kp-indices is limited in so far as they can record only such radiation that actually affects the earth: there is ample evidence that many clouds of solar gas fail to hit the earth, not only by crossing the earth's orbit somewhere else but sometimes also by shooting beyond the earth, northward or southward. The "narrow-cone" character of the solar streams that affect geomagnetism, as compared with the "wide-angle flares" of cosmic radiation emitted in rare cases from the sun, has again appeared in the cosmic-ray surge of February 23, 1956, ejected from the western edge of the solar disk.

Speaking in 1956, at the beginning of a new sunspot cycle, we may recall that the new sources of solar gas clouds, lying in the spot zones appearing in higher heliographic latitudes, fail more often to affect the earth than do those sources nearer the solar equator. That the first magnetic storms of the new cycle have their origin in the zones of sunspots seems to be proved by their (weak) recurrence tendency, with an interval of 27.75 days (average for three years after the minimum), corresponding to the rotation period for about 20° heliographic latitude, as compared with the shorter period of 27.03 days (corresponding to the solar equator) in the three years before the minimum.

For the International Geophysical Year, additional indices are planned to subdivide the 3-hour-interval into quarter-hours centered at 00, 15, 30, and 45 minutes of each hour, to agree with the proposed auroral observation schedule. They will presumably be known as Q-indices (quarter-hourly). Only stations poleward of about 58° geomagnetic latitude will measure them. While the K-index measures ranges only, the Q-index will be based on maximal deviations from normal, quiet-day curves, in order to take full account of SD (the solar daily variation on disturbed days), which is so pronounced in polar regions. The following scale will be proposed:

$$Q = 0\,1\,2\,3\,4\,5\,6\,7\,8\,9;\ T = 10;\ E = 11,$$

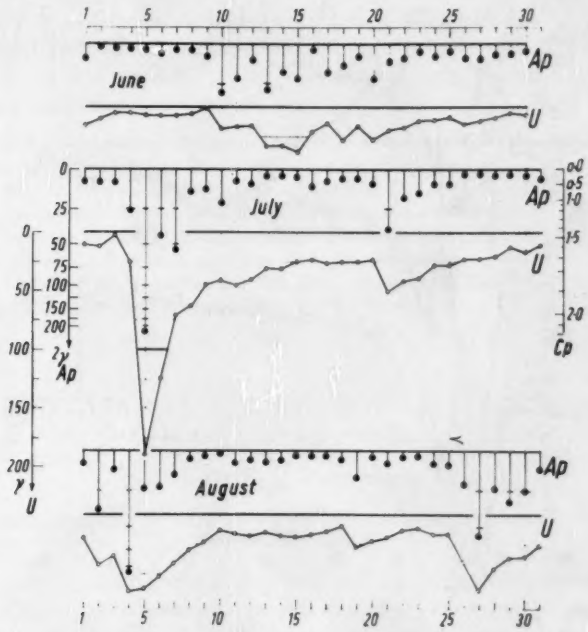
if the deviation exceeds the following number of gammas (1 gamma $= 10^{-5}$ cgs):

$$\Delta = 0 \ldots 10 \ldots 20 \ldots 40 \ldots 70 \ldots 120 \ldots 330 \ldots 500 \ldots 800 \ldots 1200 \ldots 1800\gamma.$$

In contrast to K-scaling, there will be only one and the same Q-scale for all stations. This makes it necessary to introduce the additional steps $Q = 10$ and 11, which actually occur, e.g., at College, Alaska; for convenience in tables they will be indicated by one-letter symbols $T = 10$, $E = 11$.

In Q-scaling, the vertical component Z will be disregarded, because in subauroral latitudes it is much affected by the field of the long line currents of the SD-system along the auroral zone. Thus a strong auroral current near Iceland would raise Q in Central Europe to 2 or 3, while the ionosphere overhead might be less disturbed. Restriction to the horizontal geomagnetic components has the advantage that Q then refers mainly to the ionospheric electric currents overhead of the station. With the usually adopted ratio of exterior and interior parts in the horizontal magnetic components (about 6 to 4), one may say that every gamma in Δ is the effect of an equivalent horizontal current sheet in the ionosphere with the current density 1 ampere per kilometer. $Q = 11$ corresponds to ionospheric current surface densities of at least 1,800 amperes per kilometer or more.

The first studies on Q-indices[1] made on magnetograms of Meanook, Sodankylä, College, Sitka, and Eskdalemuir show the great daily variation in geomagnetic activity just as in the K-index. At Meanook and Sodankylä, the Q-indices before local noon lie systematically lower by one to three units than before local midnight. In the two months May and June, 1933, at Meanook, less than 1 per cent of the Q-indices reached $Q = 6$ or more before noon, while, before midnight, nearly 14 per cent reached $Q = 6$ or more. (See Figures 2(a) and 2(b).)



Daily Mean Ap and Equatorial Ring-Current-Intensity U

1941, June - August

Fig. 3

The Q-index will help to express the individual changes of geomagnetic activity at each observatory. It is hoped that all polar magnetic observatories will measure Q-indices—in order to enable them to do so, the scheme providing 96 indices per day will be kept as simple and time-saving as possible. It will then be possible to adapt the standardizing process devised for Kp to arrive at some planetary index Qp. But Q and Qp will always remain supplements to, not substitutes for, K and Kp. These proposals, which will be worked out in detail, will have to be sanctioned by the IAGA (International Association of Geomagnetism and Aeronomy).

Another "corpuscular" feature of geomagnetic time fluctuations which can be used to derive characteristics of the magnetic field around the earth in general is the post-perturbation as an indicator for the intensity of the "Equatorial Ring

Current" (*ERC*). While the physics of this phenomenon remains rather obscure, its effects are clear enough. W. Kertz[2] is engaged in studies aiming to derive a continuous measure of *ERC*, which will be useful in connection with studies on cosmic-ray fluctuations. (See Figures 3 and 4.)

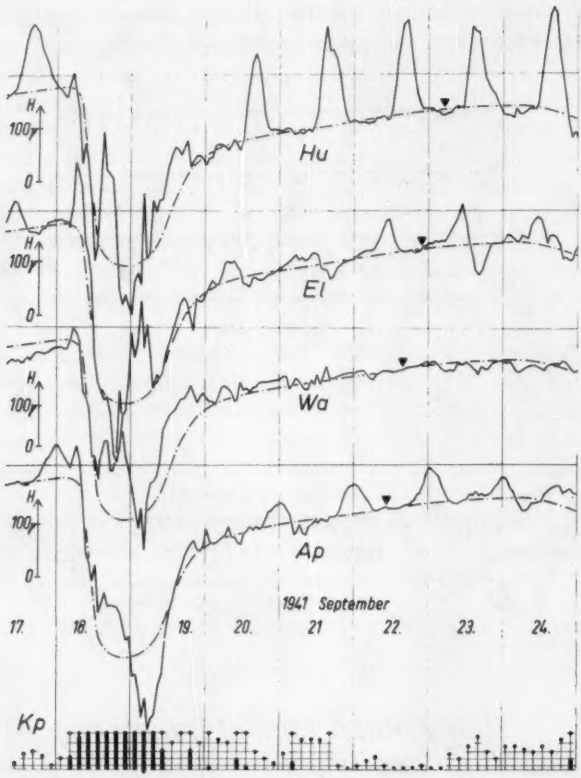


Fig. 4.—Current hourly measure of H for Huancayo, Elizabethville, Waterloo, Apis (after Kertz).

As to the wave radiation W of the sun, the daily amplitudes of $Sq + L$ at equatorial stations have been shown to be useful for the derivation of a measure for W. A uniform series, 1922–1947, has been provided from Huancayo H, of *daily* values (except for days of strong disturbance) for the intensity of solar wave radiation.[3]

Monthly indices of solar activity similar to the geomagnetic W-measure have been based on the mean noon critical frequencies of the F2-layer by C. W. Allen[4] and C. M. Minnis.[5] An attempt is now being made to arrive at *daily* values from $fF2$ noon values at Slough, in the same manner as that devised for the W-measure.

The solar-flare effects recorded in geomagnetism may be used to measure the intensity of the solar flare; they are, however, only big enough for the stronger flares. Whether the pulsations, to be studied intensely during the International Geo-

physical year, can disclose some new aspect of the solar-terrestrial relationships remains to be seen.

Finally, a plea is made to geomagnetic observatories to speed up the publication of their hourly data on magnetic components. While the simple *K*-scaling proceeds reasonably fast, it takes, in some cases, years of brooding about base-line values to produce the hourly means. Perhaps it might not be impossible to imitate the ionospheric observatories, which publish their complicated values month by month, and to publish monthly at least hourly means for the horizontal magnetic intensity *H* at tropical stations (within, say, 30° latitudes), which might be utilized in the International Geophysical Year for the computation of daily values of the *W*-measure and the *ERC*. The cessation of hourly *H*-values for Huancayo since 1948 is a particular loss to geophysics.

[1] J. Bartels and N. Fukushima, *Abhandl. Akad. Wiss. Göttingen, Math.-physik. Kl.*, 1956. The technique of scaling indices *K* and *A* is described in the *International Geophysical Year Instruction Manual*, No. III, "Geomagnetism," part 1. London, 1956.

[2] W. Kertz, *Vorträge und Berichte, Arbeitsgemeinschaft Ionosphäre und deutscher URSI-Ausschuss, Kleinheubach*, **1954** (Darmstadt: Fernmeldetechnisches Zentralamt, 1954), p. 155.

[3] J. Bartels, *Terrestrial Magnetism and Atm. Elec.*, **51**, 181, 1946.

[4] C. W. Allen, *Terrestrial Magnetism and Atm. Elec.*, **51**, 1, 1946; **53**, 43, 1948.

[5] C. M. Minnis, *J. Atm. and Terrestrial Phys.*, **7**, 310, 1955.

OBSERVATIONAL AND THEORETICAL ASPECTS OF MAGNETIC AND IONOSPHERIC STORMS

BY E. H. VESTINE

CARNEGIE INSTITUTION OF WASHINGTON, DEPARTMENT OF TERRESTRIAL MAGNETISM, WASHINGTON, D.C.

1. *Introduction.*—The geomagnetic field and the ionosphere undergo considerable changes during occasional periods of a day or so, known as "times of storm." The changes in the geomagnetic field seldom exceed 5 per cent of the whole earth's field, but during a magnetic storm the effect upon the ionosphere is often the more noteworthy, and occasionally even catastrophic, as its upper part seems actually to be blown off of the earth. There is little doubt that the sun is primarily responsible for the energy changes, as in the case of almost all other manifestations of transient energy supply on the earth. But at the present time one cannot claim to know how the sun brings about the generation of geomagnetic and ionospheric storms.

The present paper aims to describe a few of the significant features of geomagnetic and ionospheric storms and to tell something about attempts that have been made to explain these features.

2. *Representative Characteristics of Geomagnetic and Ionospheric Storms.*—Birkeland, Chapman, and others have developed methods for deriving tentative electric current systems which could reproduce the geomagnetic storm changes at ground level, when it is assumed that the electric currents flow within the ionosphere.[1] A geomagnetic storm comprises characteristic field changes in time-sequential phases usually conveniently described as the "initial phase," the "main phase," and the "recovery phase." Figures 1, 2, and 3 show more or less typical

current systems which in sequence indicate the mean hourly ionospheric electric current systems which would have magnetic fields at ground level similar to those actually observed at the fifty or more magnetic observatories which continuously record the magnetic field changes at widely distributed points over the earth's surface. Geomagnetic data at ground level were used to derive the current systems, shown, which are for the storm of April 30–May 1, 1933, of the Second International Polar Year, 1932–1933, which is the only year in which, through international cooperation, there were sufficiently large numbers of stations in polar regions to derive these current systems. In the figures the earth is viewed from above the geomagnetic north pole in northern Greenland as center of the projection, and the outer circle represents the geomagnetic equator. The current systems are supposed fixed relative to the sun, and the earth revolves within them; a total of 100,000 amperes flows between successive full-drawn current lines. The current systems wax and wane in intensity with time, so that pulses in magnetic field of several hours' to several days' duration may be experienced, at a point on the earth's surface. There also will be changes in magnetic field due to rotation of the station with the earth, and therefore relative to the current systems, as they develop in sequence. The strongest current flows (as an electrojet) near the auroral zone, where aurorae are seen most frequently and with highest intensity, on the dark side of the earth. This electrojet is directed from east to west on the night side during the main and recovery phases and may include a flow of current as great as 5,000,000 amperes on rare occasions. A smaller electrojet flows eastward on the day side. The return circuits constitute flows across the polar cap and within two oppositely directed low-latitude circuits. In low latitudes the flow of current is nearly perpendicular to the magnetic field of the earth and is mainly eastward during the initial phase and westward during the main and recovery phases. The initial phase may last about an hour or less, and the maximum intensity of the main phase is usually reached in about 16–20 hours after the beginning of the storm. The current system of the recovery phase may decay over a period of several days to even weeks.

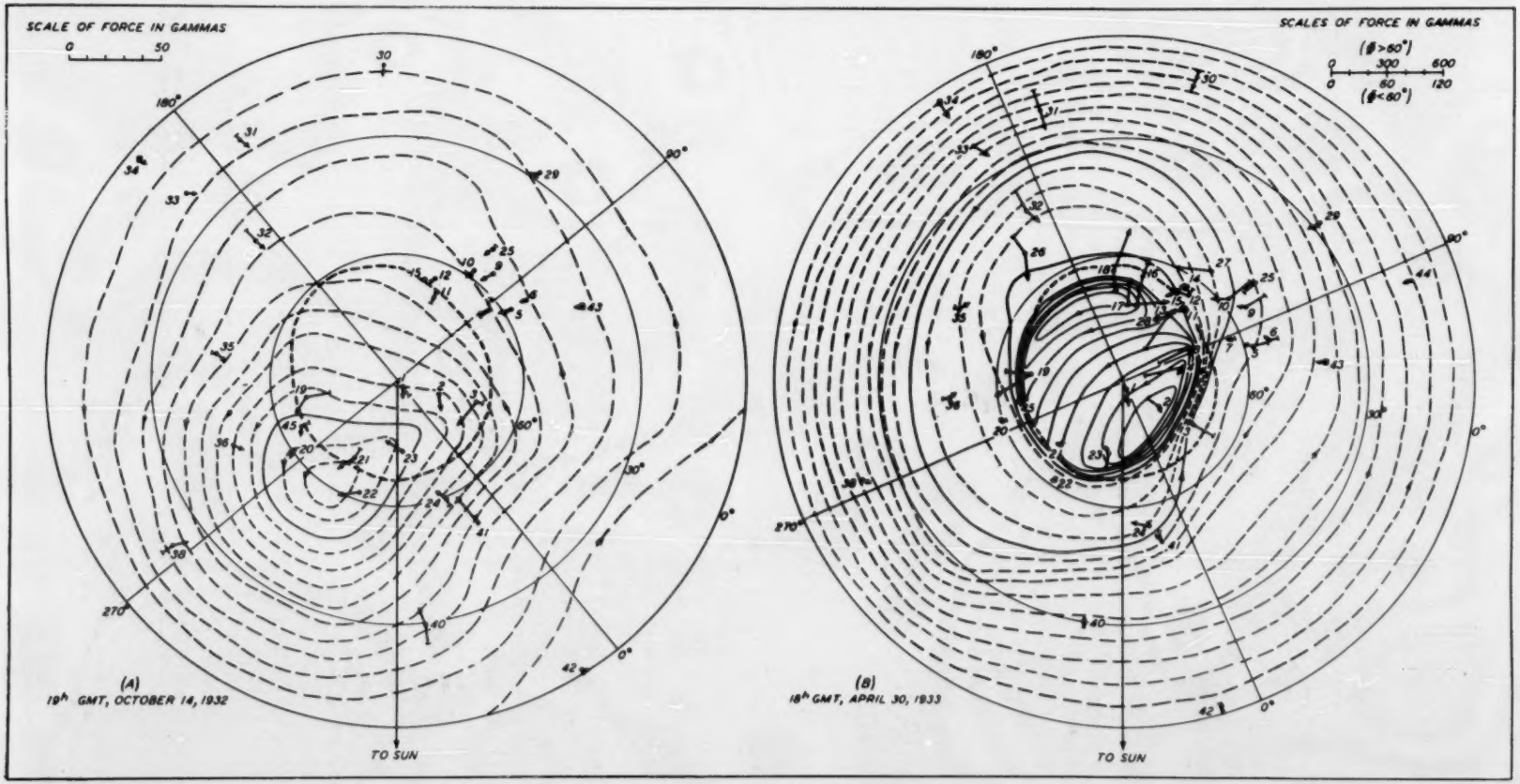
Electric forces originating in the auroral zones, northern and southern, appear to afford the main driving forces in low latitudes.[2, 3] The associated electric fields, according to Martyn, may also extend to the higher, more rarefied regions of the upper atmosphere, of which a significant part is ionized to form the F-region of the

FIG. 1.—Mean hourly disturbance vectors and corresponding electric current systems for height 150 km for maximum of initial phase of magnetic storms; view from above geomagnetic north pole.

Horizontal force →. Vertical force ——| (vertical component positive when drawn outward from geomagnetic north pole). Average auroral zone – – –. Stations missing data ■. (100,000 amperes flows between successive full-drawn current lines.)

STATIONS

1 Thule	*10* Lovö	*19* Point Barrow	*28* Sveagruvan	*37* Apia
2 Godhavn	*11* Tromsö	*20* Sitka	*29* Helwan	*38* Honolulu
3 Juliannehaab	*12* Sodankylä	*21* Fort Rae	*30* Alibag	*39* Tucson
4 Scoresby Sund	*13* Bear Island	*22* Meanook	*31* Kuyper	*40* Teoloyucan
5 Eskdalemuir	*14* Kandalakscha	*23* Chesterfield Inlet	*32* Watheroo	*41* Cheltenham
6 Greenwich	*15* Petsamo	*24* Agincourt	*33* Lukiapang	*42* Huancayo
7 Lerwick	*16* Matotchkin Shar	*25* Swider	*34* Antipolo	*43* San Fernando
8 Jan Mayen	*17* Franz Josef Land	*26* Yakhoutsk	*35* Toolangi	*44* Elizabethville
9 Rude Skov	*18* Dickson	*27* Sloutzk	*36* Christchurch	*45* College Fairbanks

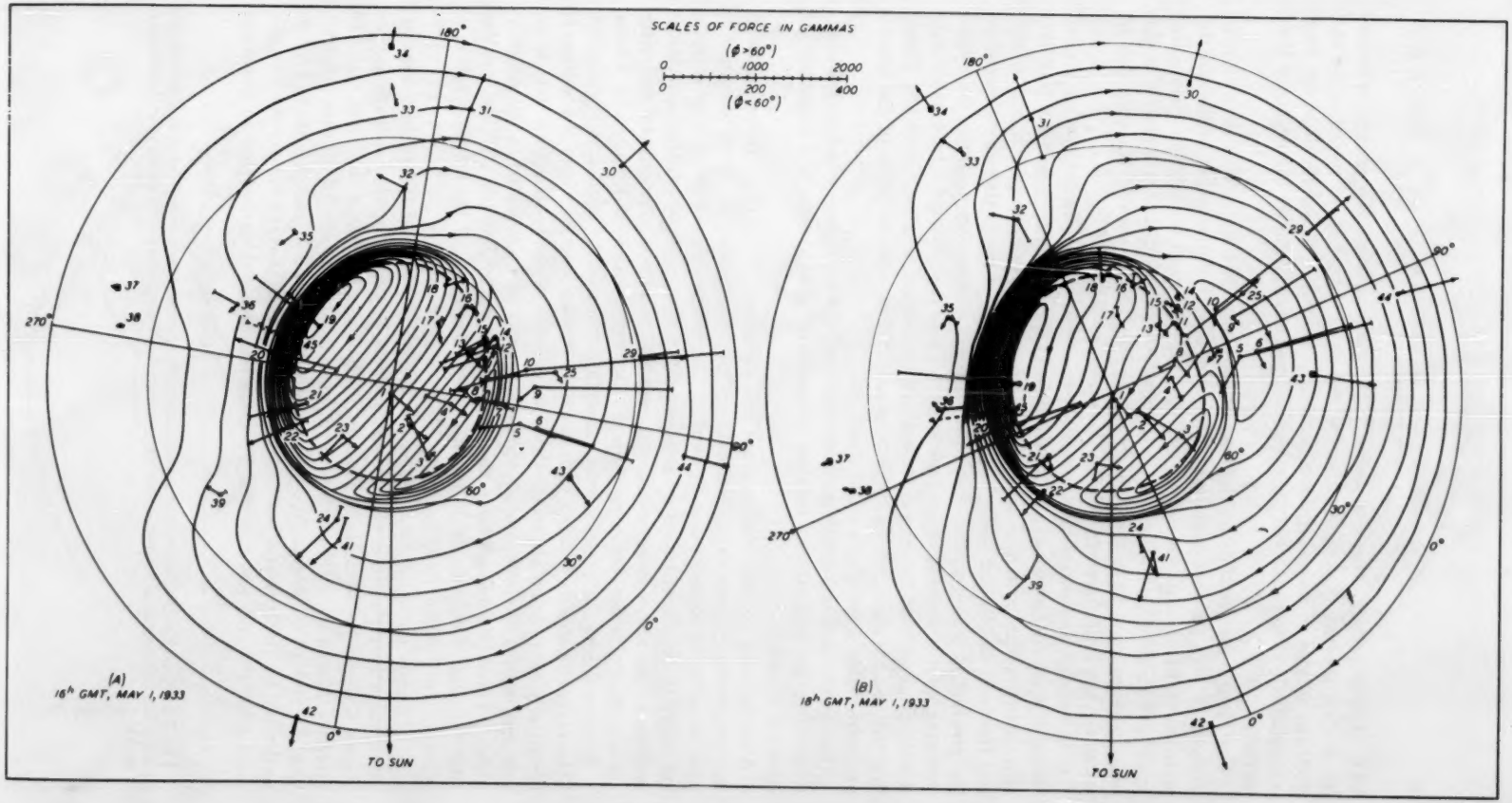

SCALE OF FORCE IN GAMMAS
0
50
(A)
19h GMT, OCTOBER 14, 1932
TO SUN
SCALES OF FORCE IN GAMMAS
(φ>60°)
0
300
600
0
60
120
(φ<60°)
(B)
18h GMT, APRIL 30, 1933
TO SUN

ionosphere at heights of 300–500 km. When these electric fields act perpendicular to the geomagnetic field in low latitudes, they may lift ion and electron concentrations upward and laterally. In an atmosphere diminishing in pressure exponentially with height, the lifting of electron clouds upward will reduce the electron concentration and hence lower the critical frequency at which radio waves will penetrate the distributions of electrons. This critical frequency is measured nearly continuously at about 50 ionospheric stations distributed over the earth.

Figure 4 illustrates the departures from normal noted in the critical frequencies for the F-region (ordinary wave), during the initial, main, and recovery phases of several averaged magnetic storms, according to Obayashi.[4] The representations are for the northern hemisphere, according to geographic coordinates, which for present purposes differ insignificantly from the geomagnetic coordinates used for the current systems just noted. The departures from median values of critical frequency are shown in megacycles per second. As in Figures 2 and 3, for electric currents, the low-latitude patterns which tended to remain more or less fixed about the 6:00 A.M. meridian or relative to the sun during the main phase tend later to follow the earth somewhat (though with very considerable slippage) during the recovery phase. In polar regions the electrojets seem to show an opposite tendency, but it is not clear that the ionospheric patterns move likewise. The average values K_p indicated by Obayashi are indices providing an indication of the changes in strength of the polar electrojets with time.

It may be instructive at this stage to consider a qualitative explanation of the possible corresponding motions in current flow and ionospheric patterns that are indicated in Figures 1–4. Martyn[3–5] has pointed out the importance of electric fields in effecting vertical and lateral motions in the F-region and has indicated that the electric fields originating in the polar electrojets may profoundly affect the F-region in low latitudes, together with the fields there of electrostatic charge distributions driving the E-region electric currents responsible for the ordinary magnetically quiet-day diurnal variation. Using his suggestion, it would appear tentatively that an electric field E directed poleward at the auroral zone, on the 6:00 A.M. meridian, should through $E \times H$ forces drive ions and electrons alike with about the same velocity to the west in the F-region and would give currents flowing to the east in the E-region, where, due to the higher collisional frequency for ions, the relative mobility of the electrons will dominate. Since the observed currents flow to the west, an electric line doublet directed away from the pole at E-region, so that that field is poleward in the F-region above, may at any rate drive auroral rays in the F-region above to the west, as observed by Meinel[6] and others. In low latitudes the electric field, if due to polar doublets or otherwise, should be directed downward, and on the morning and day side also to the east, so that the F-region ionization may be driven both upward and eastward in accord with Obayashi's data. In fact, the currents that must then be driven in the E-region would even flow in the right direction as Hall currents, perpendicular to the geomagnetic

FIG. 2.—Mean hourly disturbance vectors and corresponding electric current systems for height 150 km for main phase of magnetic storms; view from above geomagnetic north pole (legend as in Fig. 1).


SCALES OF FORCE IN GAMMAS
(φ>60°)
0
1000
2000
0
200
400
(φ<60°)
180°
270°
90°
0°
60°
30°
TO SUN
(A)
16h GMT, MAY 1, 1933
(B)
16h GMT, MAY 1, 1933
TO SUN

field. Unfortunately, there is no present indication that the polar electrojets are electric generators, acting either by the dynamo action of winds or by the electrical actions of incoming solar particles; moreover, it is difficult to see why polarizing effects in equatorial regions should not quickly nullify the proposed Hall currents.

The electric conductivity at the magnetic equator on the day side may be relatively very high for eastward-directed electric fields,[7] which seems to explain the presence of locally intense electric currents there during the sudden commencement and initial phase of magnetic storms[8-10]; electric fields originating in the polar regions may drive these electric currents. Polar electric currents due to initial electric fields, have, in fact, been deduced by Nagata.[11]

It is usually considered more plausible to regard the westward-flowing atmospheric current circulation as a representation of the field of an equatorial ring current encircling the earth at a distance of a few earth radii.[1] In fact, this is a central theme of the Chapman-Ferraro-Martyn theory of magnetic storms, based on the creation of a ring current from the action of the earth's magnetic field upon incoming solar streams of ionized particles. These particles are subsequently supposed to leak off from the ring current along the lines of magnetic force and thence into the polar regions, where for reasons as yet quite obscure some of the sources of electric fields noted above might result.

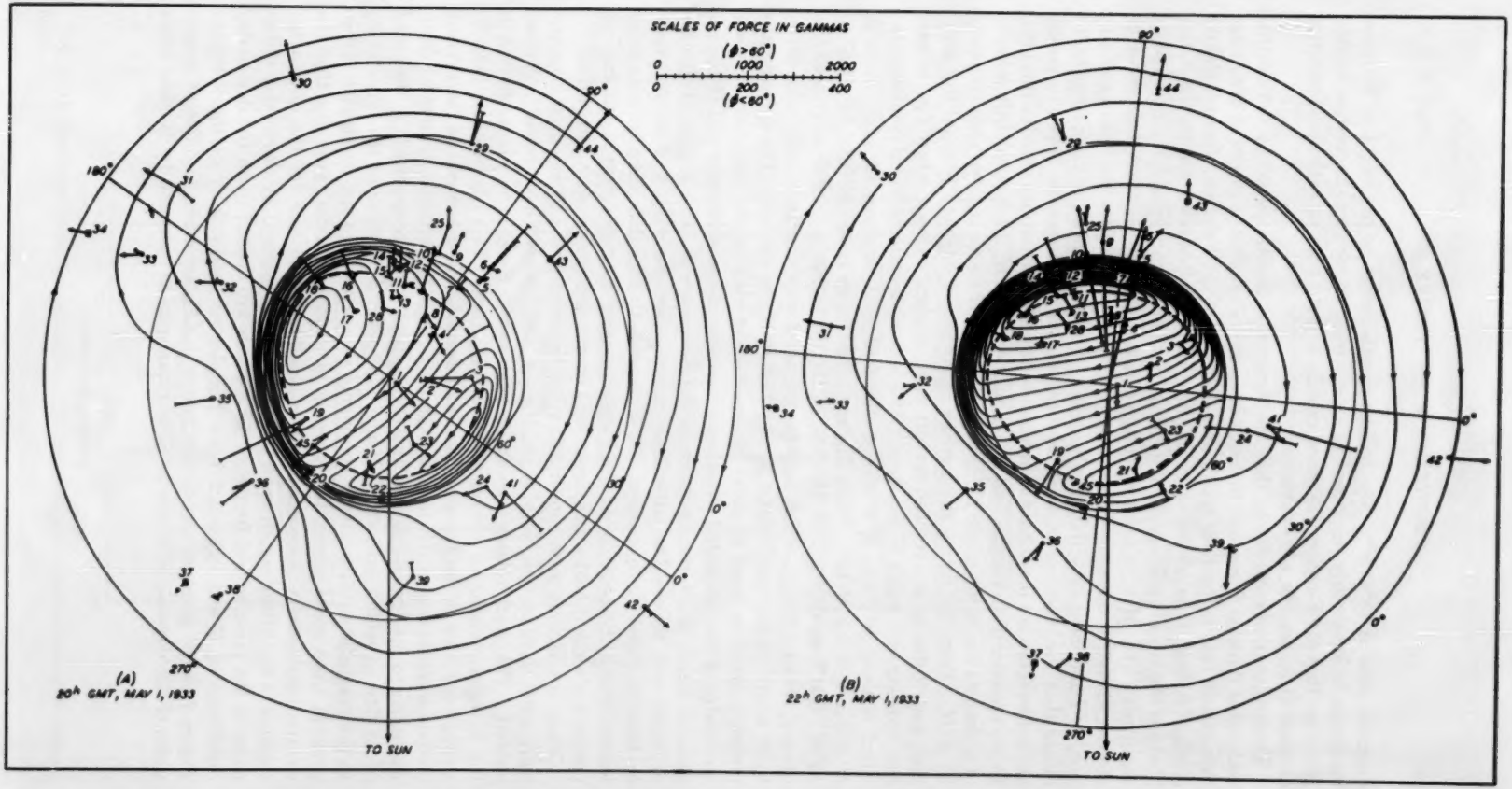
Observational data for both geomagnetism and the ionosphere which might be attributed to the ring current because independent of local time and longitude are shown in Figure 5.

A modernized version, due to some new additions, of the older Birkeland-Störmer corpuscular theory of magnetic disturbance and aurora has been brought forward recently by Bennett and Hulburt.[12] According to their ideas, solar streams are likely to be concentrated by self-focusing action, after which they arrive and penetrate the atmosphere in auroral regions. Another solar-stream and ring-current theory, in which electric fields are important, has been discussed by Alfvén.

Another theory, at present most strongly advocated by Wulf and Nicholson,[13] links magnetic storms to wind storms within the ionosphere, where the motion of electrically conducting air across the lines of force of the geomagnetic field is the principal source of electromotive forces, especially to be expected in the polar regions. These possibilities have also been discussed by Fukushima[14] and by Vestine,[15] though it now appears that the latter's discussion of the dynamo action of winds in the F-region must be modified.

3. *Changes during a Severe Magnetic Storm.*—Occasional magnetic storms are accompanied by magnetic changes as much as 10 times those of average storms. Figure 6 shows the ionospheric changes in the F-region during the great magnetic storm of March 24, 1940 at Huancayo, Peru (south latitude 12°.0, east longitude 284°.7), and at Watheroo, Australia (south latitude 30°.3, east longitude 115°.9), as derived by Berkner and Seaton.[16] The initial sudden commencement was sometime after 13^h and was followed by a much larger sudden increase in horizontal

FIG. 3.—Mean hourly disturbance vectors and corresponding electric current systems for height 150 km for main phase of magnetic storms; view from above geomagnetic north pole (legend as in FIG. 1).


SCALES OF FORCE IN GAMMAS
(φ>60°)
0
1000
2000
0
200
400
(φ<60°)
(A)
20h GMT, MAY 1, 1933
TO SUN
(B)
22h GMT, MAY 1, 1933
TO SUN

magnetic field at $15^h44.8^m$. The broken lines below show the height of F-region maximum in ion density, expressed as a ratio from division by the median height. It is clear that the F-region was raised rapidly to great heights, with accompanying decrease in ion density, at both stations.

The explanation seems to be that the F-region is driven upward by the electrostatic field directed from west to east within it. The obvious means of obtaining this electric field is from the separation of charge in the E-region, where it should produce strongly augmented eastward-directed electric currents during the initial phase of the storm as observed, and the motions at Huancayo and Watheroo seem compatible with Martyn's concept of polar electrojets. This requires an excess of positive charge to the west of Huancayo and negative charge to the east of this station, probably near or within the E-region. Some rough calculations seem to indicate that the observed motions in the F-region and magnetic field changes at ground level are not incompatible in magnitude with expectations based on this idea. In the presence of an electric field E and magnetic field H, the vertical drift velocity of a charged ion in the upper F-region due to $\boldsymbol{E} \times \boldsymbol{H}$ forces is given approximately by E/H. From the figure, the increase in horizontal intensity is about 600 gammas, and, if due to an overhead current sheet, is equal to $2\pi i$, where i is the current crossing unit width of the sheet. If k is the integrated vertical conductivity, which is about 10^{-7} cgs, $E = 600 \times 10^{-5}/6 \times 10^{-7}$, or about 10^4 cgs. Thus, if $H = 0.3$, $v = 3.3 \times 10^4$. From Figure 6, the change in height from $15^h\ 41.5^m$ to 16^h06^m GMT, or during 24.5 minutes, is, for the region of maximum critical frequency, about 720 km., which may be compared with the theoretical value given by vt, where t is the time, or $24.5 \times 60 \times 3.3 \times 10^4$, or 490 km. The agreement is probably about as good as could be expected, in view of the uncertainty in our knowledge of the integrated electric conductivity of the ionosphere and of the associated details of ion concentration of the F-region with large and rapid changes in height with time. The latter have recently been the subject of careful study by Ratcliffe[17] and his co-workers. It is plausible to infer that much remains to be learned about the behavior of the F-region during magnetic storms, but the importance of electric field actions upon the F-region appears certainly established.

These conclusions seem supported by the behavior of the F-region on magnetically quiet days as well, as Martyn, Hirono, and Maeda[18] have recently demonstrated. This work relates the electric fields in the F-region to winds blowing in the E-region, where an electrostatic field E arises. This electrostatic field E, strong near noon in the E-region, serves to close the electric circuit of the varying electric currents responsible for the solar daily magnetic variation S_q on quiet days. According to the dynamo theory,[7] this electrostatic field is proportional to the speed of air flow in middle latitudes. Hence day-to-day differences in the range of S_q at the magnetic equator should be roughly proportional to the day-to-day differences in wind speed within the E-region in middle latitudes, assuming that the electric conductivity is constant. This should also be true for the day-to-day changes in the lifting action upon the F-region, as indicated by day-to-day differences in the noon critical frequency f_0F_2, which will depend upon the day-to-day differences in E and not much upon the electric conductivity of the E-region. Figure 7 shows averaged day-to-day ranges in f_0F_2, according to different magnitudes of the day-to-day ranges in S_q at Huancayo, as measured by Bartels' index

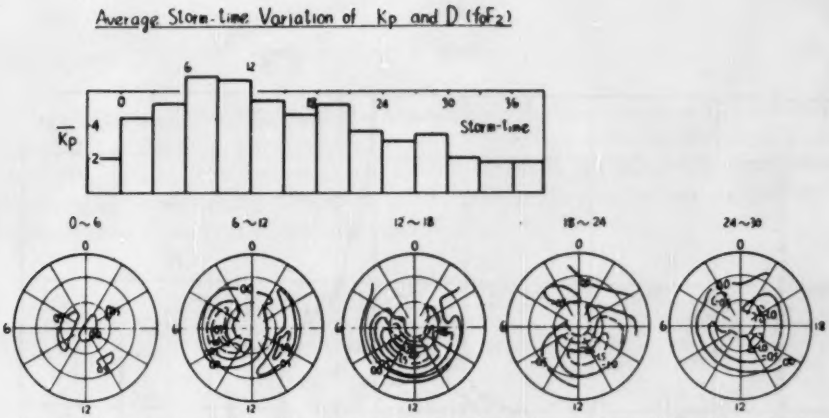


FIG. 4.—Average storm-time changes of magnetic activity K_p and the world patterns of corresponding ionospheric disturbance $D(f_0F_2)$, deviations of f_0F_2 (in mc/s).

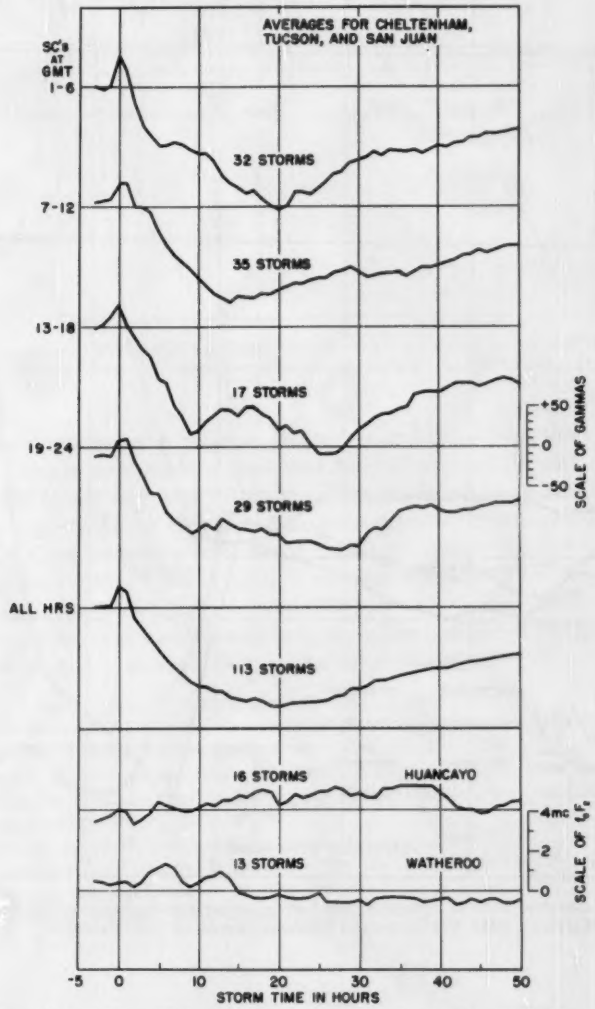


FIG. 5.—Magnetic and ionospheric storms averaged according to Greenwich times of commencement.

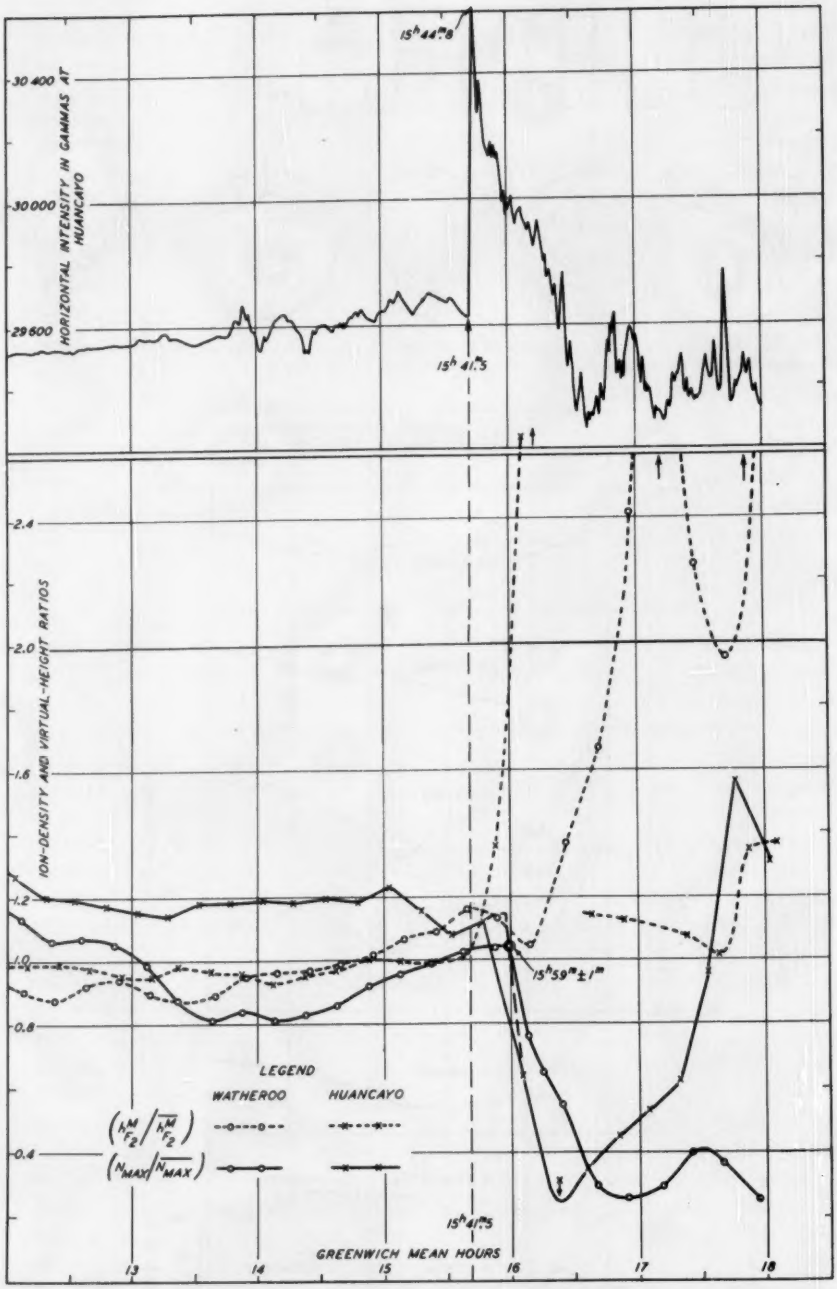


FIG. 6.—Comparative time of commencement of ionospheric disturbance coinciding with magnetic storm of March 24, 1940, Watheroo and Huancayo magnetic observatories.

of solar wave radiation W_2,[19] which is derived from the daily range in S_q, corrected for effects of the lunar daily magnetic variation; values of W_2 on disturbed days were eliminated. The regression line for differences in W_2 selected according to the sizes of corresponding differences in f_0F_2 showed that the correlation between differences in W_2 and in f_0F_2 is about -0.8.

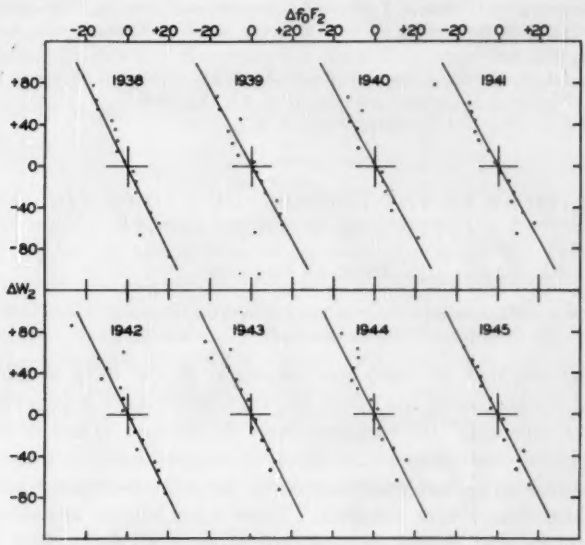


FIG. 7.—Day-to-day changes in amplitude of daily variation S_q (in terms of differences in Bartels' W_2) and F-region critical frequencies f_0F_2 at noon, Huancayo, 1938–1945. Averages according to ranges ΔW_2.

It will be noted that the slopes of the curves in Figure 7 for the various years, 1938–1945, change only slightly from year to year, so that the day-to-day differences in wind speeds near and within the E-region do not appear to change much with sunspot cycle, though they may change considerably from one day to the next. The average slope is at the rate of about -0.38 mc/unit of W_2, the unit of W_2 being 2×10^{-5} cgs. On the other hand, it was found that both f_0F_2 and W_2 increase with sunspot number, at a rate of about 0.15 mc/unit of W_2. There are also day-to-day changes in wind speed determined in similar fashion which indicate a variability, with season, which will be reported upon separately.

[1] S. Chapman and J. Bartels, *Geomagnetism* (2 vols.; Oxford, 1940).
[2] H. B. Silsbee and E. H. Vestine, *Terrestrial Magnetism and Atm. Elec.*, **47**, 195, 1942.
[3] E. H. Vestine, *Terrestrial Magnetism and Atm. Elec.*, **43**, 261, 1938.
[4] T. Obayashi, *J. Geomagnetism and Geoelectricity*, **6**, 57, 1954.
[5] J. A. Ratcliffe, *Physics of the Ionosphere*, The Physical Society, London, (1955), p. 88; also D. F. Martyn, *ibid.*, pp. 88, 163, 212, 260.
[6] A. B. Meinel and D. H. Schulte, *Astrophys. J.*, **117**, 454, 1953.
[7] W. G. Baker and D. F. Martyn, *Phil. Trans. Roy. Soc. London, A*, **246**, 281, 1953.
[8] M. Sugiura, *J. Geophys. Research*, **58**, 558, 1953.
[9] E. H. Vestine, *J. Geophys. Research*, **58**, 539, 560, 1953.

[10] S. E. Forbush and E. H. Vestine, *J. Geophys. Research*, **60**, 299, 1955.

[11] T. Nagata, *Report of Ionosphere Research in Japan*, **6**, 13, 1952; also T. Nagata and S. Abe, *Report of Ionosphere Research in Japan*, **9**, 39, 1955.

[12] W. H. Bennett and E. O. Hulburt, *Phys. Rev.*, **91**, 1562, 1953.

[13] S. B. Nicholson and O. R. Wulf, *J. Geophys. Research*, **60**, 389, 1955.

[14] N. Fukushima, *J. Fac. Sci. Univ. Tokyo*, **8**, 293, 1953.

[15] E. H. Vestine, *J. Geophys. Research*, **59**, 93, 1954.

[16] L. V. Berkner and S. L. Seaton, *Terrestrial Magnetism and Atm. Elec.*, **45**, 393, 1940.

[17] J. A. Ratcliffe, E. R. Schmerling, C. S. G. K. Setty, and J. O. Thomas, *Phil. Trans. Roy. Soc. London, A*, **248**, 609, 1955.

[18] M. Hirono and H. Maeda, *J. Geophys. Research*, **60**, 241, 1955, and references there cited.

[19] J. Bartels, *Terrestrial Magnetism and Atm. Elec.*, **51**, 181, 1946.

[20] M. J. Seaton, *J. Atm. and Terrestrial Phys.*, **8**, 122, 1956.

VARIATIONS OF THE IONOSPHERE AND OF THE SOLAR INTENSITY WITH SUNSPOTS

BY E. O. HULBURT

U.S. NATIONAL COMMITTEE FOR THE INTERNATIONAL GEOPHYSICAL YEAR
NATIONAL ACADEMY OF SCIENCES, WASHINGTON, D.C.

It is well known that the solar spectral energy in the wave lengths which are observable at the surface of the earth, i.e., from about 3000 A to several microns, is remarkably constant. Its variation over the sunspot cycles is less than 0.1 per cent and is not observable within the error of experiment. On the other hand, the energy in the very short wave lengths of the solar spectrum which cause the ionosphere does change with sunspots. These wave lengths are absorbed in the upper atmosphere and cannot be observed directly at the surface. But their changes may be inferred from the observed changes in the ionosphere. The variation of the short wave lengths is worked out over two sunspot cycles in the following paragraphs.

By means of apparatus on rockets the intensity in the far ultraviolet[1] and X-ray[2] regions of the solar spectrum practically outside the atmosphere has been measured. Also, curves have been obtained of the vertical distribution with height of the ionization in the ionosphere up to about 200 kilometers.[3] From these observations theories have been worked out for the main ionospheric regions which attributed the D-region to the Lyman-alpha line of hydrogen at 1216 A, the E-region to X-rays from about 10 to 20 A, and the F2-region to ultraviolet wave lengths in the 100–600 A interval.[4] The origin of F1 has not yet been discovered. There seems to be no doubt that F1 was caused by some wave lengths between 10 and 1000 A, but the details of its photochemistry are not known. F1 is a bump or protuberance of the main F-region and is observable only when the sun is more than about 30° above the horizon.

In any ionospheric region, let y be the electron density at a height z and zenith angle Z of the sun; let y_m be the value of y at the maximum of the y, z curve and y_0 the value of y_m when the sun is directly overhead, i.e., $Z = 0$. Further, let $\bar{y}$ be the total ionization in a vertical column through the region below y_m.

The normal E-region, which is at about 100–120 km. altitude, is the most regular of the ionospheric regions and follows solar control very closely. Parenthetically,

it should be mentioned that we are not concerned here with the sporadic E-region, which is nearly in the same place as normal E but which is produced by agencies other than solar light, probably meteors for the most part. In an earlier paper[5] it was shown that for the normal E-region during the daylight hours

$$y_m = y_0 (\cos Z)^{1/2}. \quad (1)$$

At any ionosphere station equation (1) was true within a few per cent for changes in Z throughout the day and the year, but for widely scattered stations from 80° north to 30° south latitude, y_m calculated from equation (1) for a given time varied by as much as 40 per cent. From equation (1) it followed that, within about 40 per cent,

$$y_0 = a i_0^{1/2}, \quad (2)$$

where a is a constant and i_0 is the intensity of the solar ionizing wave lengths falling vertically on top of the atmosphere which produce normal E. The square-root relation of equation (2) is in accord with a loss of E-region electrons by dissociative recombination.

In order to show the manner in which the ionosphere varied with solar activity, y_0 of E, averaged over the year for the years 1931–1954, was calculated from the data[6] of a number of ionosphere stations scattered over the earth. The values are plotted in the upper curve of Figure 1. The yearly average Wolf[7] sunspot number

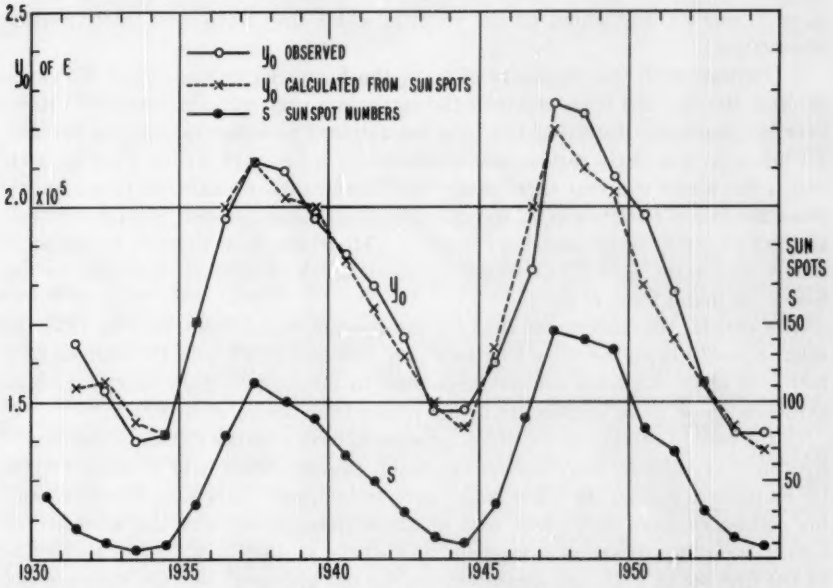


FIG. 1.—Annual mean values of Wolf sunspot number s; y_0 of E observed and calculated.

s is plotted in the lower curve of Figure 1; it is seen that the y_0 and s curves rise and fall together. A theoretical relation between y_0 and s was derived by assuming

that the ionizing radiation was proportional to the sunspot number. Then

$$i_0 = b(g + s), \tag{3}$$

where b and g are constants, g being introduced in recognition of the fact that i_0 is not zero when $s = 0$, i.e., when there are no sunspots. Then, from equations (2) and (3),

$$y_0 = f(g + s)^{1/2}, \tag{4}$$

where f is a constant which replaces a $b^{1/2}$. The dashed curve of Figure 1 was plotted from equation (4), with s from the lower curve of Figure 1 and with $f = 0.154 \times 10^5$ and $g = 78.6$. It agreed with the observed curve fairly well, showing that the Wolf sunspot number is a surprisingly good indicator of the intensity i_0 of the soft X-rays or short ultraviolet wave lengths which cause the E-region. From sunspot minimum to maximum y_0 increased by a ratio of about 1.5; hence i_0 increased by 1.5^2, or about 2.3.

As a matter of fact, in the earlier paper[5] it was shown that E-region ionization was related to sunspots by equation (4) for the years 1932–1938, and what has been added here is the fact that the same relation held for the next 16 years.

The F1-region is known, or exists, only for Z less than about 60°, because, for larger values of Z, F1 is engulfed, or covered up, by F2 and is concealed from observation. However, where it is observed, F1 obeys equation (1) approximately, and y_0 of F1 is about $2y_0$ of E. Hence for F1 equation (4) is true with $f = 0.3 \times 10^4$ and $g = 78.6$. Thus Figure 1 is a concise statement of the major features of y_m of E and F1 throughout all the years in which they have been under routine observation.

In contrast with the regularity of y_m of the E and F1 regions, y_m of F2 varies through the day and year, and over the earth, in a very complex manner.[8] However, as shown by Ratcliffe,[9] the total ionization $\bar{y}$ in a vertical column through F2 below y_m was fairly regular and decreased by a factor of 2 from 1939 to 1944, which was about one year short of the span from sunspot maximum to minimum. Since the loss of electrons in F2 is by the charge-exchange process, $\bar{y}$ is proportional to the first power of the ionizing energy i_0. Therefore, from sunspot minimum to nearly maximum, i_0 of F2 increased by about 2, which is about the same as the factor 2.3 found for i_0 of E.

The general conclusion was that for the sunspot cycles from 1931 to 1954 the solar intensity in all the wave lengths which produced E, F1, and F2 changed by a factor of about 2.3 from sunspot maximum to minimum. According to present evidence, these wave lengths are in the region from about 10 to 800 A.

In the case of the D-region, which appears to be due to the Lyman-alpha line of hydrogen at 1216 A, there are no ionization measurements over a sunspot cycle by ionosphere stations or other radio-wave techniques. However, Friedman and his colleagues have made four sets of direct measurements of the intensity of Lyman-alpha by means of equipment on rockets, in 1949, 1952, 1954, and 1955. In the first set in 1949, the measurements did not appear to be very accurate but indicated a value for the intensity of Lyman-alpha of 1–10 ergs/cm^2/sec; from Figure 1 s was 140 in that year. In 1952, and 1954, Lyman-alpha was 0.1 and 1 erg/cm^2/sec and s was 30 and 8, respectively. In 1955, s had increased to more

than 40, and on three rocket flights in that year Lyman-alpha was 4, 5.7, and 9.2 ergs/cm^2/sec. The general conclusion at present is that the intensity of Lyman-alpha increases and decreases erratically with the sunspot number.

Summary.—From the observed changes in the ionosphere over two solar cycles from 1932 to 1954 the solar intensity in the wave lengths which cause the E-, F1-, and F2-regions was calculated to increase by a factor of about 2.3 from sunspot minimum to maximum. The cause of E is attributed probably to X-rays of wave lengths 10–20 A and of F2 to short ultraviolet wave lengths in the region 100–600 A.

[1] F. S. Johnson, J. D. Purcell, R. Tousey, and N. Wilson, "The Ultraviolet Spectrum of the Sun," in *Rocket Exploration of the Upper Atmosphere*, ed. R. L. F. Boyd and M. J. Seaton (New York: Interscience Publishers, 1954), pp. 279–288. (A survey paper with bibliography.)

[2] H. Friedman, "The Solar Spectrum below 2000 Angstroms," *Ann. Geophys.*, **11**, 174–180, 1955. (A survey paper with bibliography.)

[3] J. C. Seddon, A. D. Pickar, and J. E. Jackson, "Continuous Electron Density Measurements up to 200 Km," *J. Geophys. Research*, **59**, 513–524, 1954.

[4] R. J. Havens, H. Friedman, and E. O. Hulburt, "The Ionospheric F2 Region," in *Proceedings of Conference on the Physics of the Ionosphere, Cambridge, England, September 1954* (London: Physical Society, 1955), pp. 237–244.

[5] E. O. Hulburt, "The E Region of the Ionosphere," *Phys. Rev.*, **55**, 639–645, 1939.

[6] *Ionosphere Data* (Central Radio Propagation Laboratory, U.S. National Bureau of Standards).

[7] "Wolf's Sunspot Numbers, Annual Mean," in *Smithsonian Physical Tables* (9th rev. ed., 1954), p. 727, Table 824.

[8] D. F. Martyn, "Geomagnetic Anomalies of the F2 Region and Their Interpretation," in *Proceedings of Conference on the Physics of the Ionosphere, Cambridge, September 1954* (London: Physical Society, 1955) pp. 260–264.

[9] J. A. Ratcliffe, "Some Regularities in the F2 Region of the Ionosphere," *J. Geophys. Research*, **56**, 487–507, 1951.

SOLAR RADIATION AND THE LOWER ATMOSPHERE

BY SIGMUND FRITZ

U.S. WEATHER BUREAU, WASHINGTON, D.C.

Introduction.—The sun may conceivably affect the major circulation of the lower atmosphere in several ways. One way is through the influence of variable solar emissions; meteorologists have long speculated about the possible effects of irregular solar emissions upon the general circulation. "Abnormal" wave or particle emissions from the sun, absorbed either high in the atmosphere or even at the ground, have been studied in many ways in connection with meteorological variables of the lower atmosphere or troposphere. A less glamorous solar influence is that of the nonvariable sun. In this case the excess net heating in the tropics and summer hemisphere, on the one hand, and the net radiative cooling in much of the winter hemisphere, on the other hand, coupled with the earth's rotation, produce the basic general circulation pattern. On this meteorologists agree; regarding the details of the general circulation and how they come into existence, there is much less agreement.

The Variable Sun (Direct Heating).—Let us consider first the basis for the so-called "solar-weather relations" which involve the variable sun. Interesting

relations have been suggested from time to time, of which some recent ones are those of Duell and Duell,[1] Wexler,[2] and Farthing[3]; many others exist. In general, each of these studies has used some parameter which supposedly represented an undefined wave or particle emission from the sun. This solar parameter was then supposed to be related to some meteorological feature of the lower atmosphere. For example, the Duells related magnetic character figure C to atmospheric sea-level pressure near Iceland. Wexler suggested a relation between sunspots and world-wide January surface temperature and pressure distribution. Farthing used coronal measurements to forecast temperature and precipitation at Kansas City. These studies have all been statistical in nature, and it was usually not clear whether the results were statistically significant or not. No plausible physical explanation has yet been accepted.

Before one can determine the physical basis for claims about "solar-weather" relations, it will be necessary to examine the spectral emission of both the quiet and the disturbed sun. If we take the energy of the quiet sun in a 1 A wavelength interval near 5000 A as unity, we can express the energy at other wave lengths as a fraction of this unit. The solid line in Figure 1 shows such smoothed data from Johnson[4] for $\lambda > 2200$ A, from Newell[5] for $\lambda < 2400$ A, and from Hulburt[6]

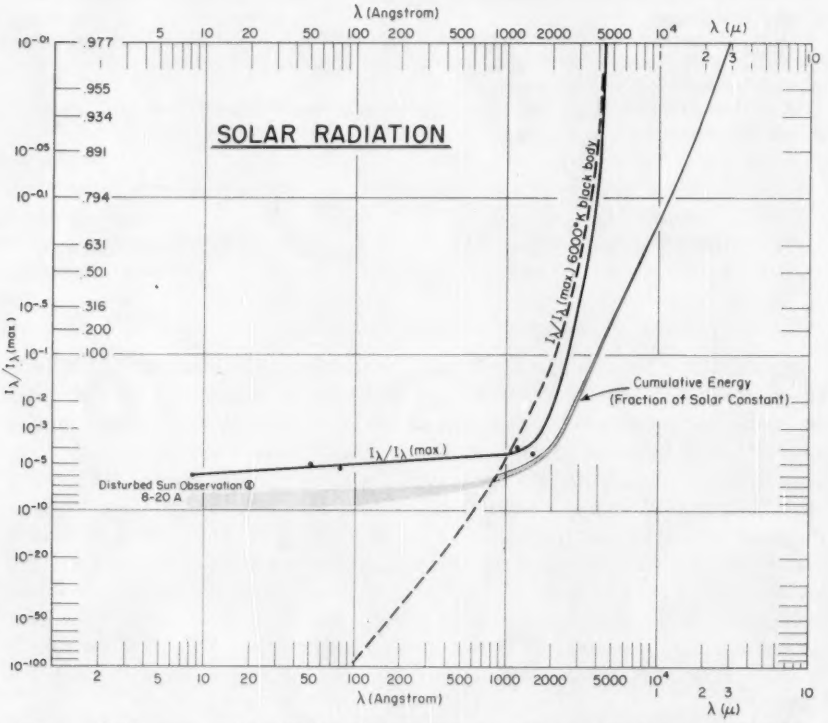


FIG. 1.—*Solid curve:* solar radiation (smoothed) from quiet sun. *Dashed curve:* radiation from a black body at 6,000° K. *Stippled curve:* cumulative solar energy, $P(\lambda)$, below wave length, λ. Symbol ⊗: cumulative energy from disturbed sun.

for $\lambda < 100$ A. For comparison, the energy in similar units for a black body of 6,000° K. is shown by the dashed curve. We note that the solar energy is about equal to that of a 6,000° K. black body near 5000 A. The solar energy falls below the black-body curve until about 1000 A is reached, although this may not be so for individual spectral lines. Then apparently the hot corona becomes more important, and by the time the X-ray region near 100 A is reached, the sun's energy is larger than that of the 6,000° K. black body by a factor of about 10^{100}. Nevertheless, the solar energy in a band of the X-ray region is still about 10^{-6} of the energy in a similar band in the visible spectrum.

The curves just mentioned give the relative spectral energy. The remaining, broadened curve in Figure 1 shows the amounts of energy, $P(\lambda)$, in the solar spectrum below any given wave length, λ, on the assumption that little or no energy exists in the solar spectrum below 6 A (Newell[5]). This energy has been expressed in units of the solar constant, taken to be 2 ly/min or 1.35×10^6 ergs/cm^2/sec, and is summarized approximately in Table 1.

TABLE 1

FRACTION OF SOLAR ENERGY, $P(\lambda)$, BELOW WAVE LENGTH λ (Unit Is Solar Constant)

λ (A)	10	100	1000	2000	3000	4000
$P(\lambda)$	10^{-10}	10^{-8}	10^{-6}	10^{-4}	10^{-2}	10^{-1}

Thus we see that rather small amounts of energy reach the outer limits of our atmosphere in the X-ray and far-ultraviolet regions of the spectrum. A few rocket measurements (Hulburt[6]) indicate that, in the X-ray region, increases of a hundred fold have occurred during disturbed sun conditions. Such an increase is indicated in Figure 1. Thus the total energy below 20 A is still only 10^{-8} of the solar constant even during disturbed sun conditions, although, of course, additional measurements may later indicate some further increase in this value.

Aside from the rocket X-ray measurements, direct spectral measurements of the emission of the disturbed sun have apparently not been made for $\lambda < 3200$ A. At $\lambda = 3200$ A, the measurements of Pettit[7] suggest a variation during the sunspot cycle. Pettit's measurements indicated that the solar intensity at 3200 A was about 1.5 times greater at sunspot maximum than at sunspot minimum. He himself discounted the variations as being too large but apparently felt that some variation at 3200 A existed in relation to the sunspot cycle.

What the magnitude of the variations is in the spectral regions between 20 A and 3200 A can at present only be estimated from indirect evidence. Lyman-alpha is expected to vary because of the observed H(α) variations in red light. Mitra[8] suggests that the variations during solar flares in L(α) cause sudden ionospheric disturbances (S.I.D.). During an intense flare, Mitra estimates an increased emission by a factor of 90 at $\lambda = 1216$ A. But Wulf and Deming[9] have suggested that emissions in the region between 2300 and 2800 A might be an important cause of S.I.D. How much of an increase, if any, might occur in these wave lengths from a disturbed sun is not known.

Of importance from the meteorological point of view is the region in the atmosphere where solar energy is absorbed. The heights at which absorption occurs

are shown in Figure 2, in which the spectral absorptions are taken from De Jager[10]; the temperature, pressure, and density data are summarized from Hulburt.[6] Just under the wave-length scale have been added the logarithms of $P(\lambda)$ taken from Figure 1. Thus in the region above 120 km., where the mass of the earth's atmosphere is about 10^{-8} of its total mass, less than 10^{-5} of the solar constant is absorbed. In the region above 60 km., with a mass of about 10^{-4} of the total mass, about 10^{-4} of the solar constant is absorbed. From 30 to 60 km. nearly 1 per cent of the solar constant is absorbed; this, too, would be the region where, if they exist,

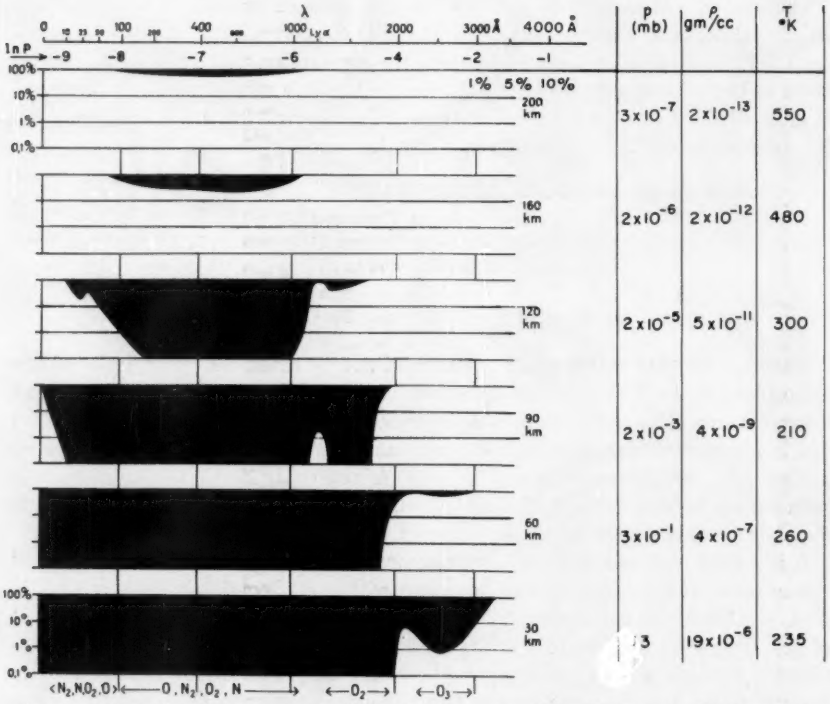


FIG. 2.—Absorption of solar energy in the earth's atmosphere (after De Jager). The base of the black areas shows the percentage of solar energy transmitted as a function of wave length at various heights. The pressure, density, and temperature are shown on the right (after Hulburt).

large increases in emission between 2000 and 3000 A (Wulf and Deming[9]) would make themselves felt. Allowing for a hundred-fold or even a thousand-fold increase in the X-ray region and/or in the ultraviolet up to $L(\alpha)$, we are faced with the question as to whether 10^{-3} of the solar constant absorbed by the upper 10^{-6} of the mass of the earth's atmosphere can have significant consequences in the earth's lower atmosphere. Most meteorologists would probably decide in the negative; but an absolute proof is certainly difficult.

Particles.—The irregular emissions from the sun include particles in addition to electromagnetic radiation. The kinetic energy of the particles which normally

impinge on the earth's atmosphere represents about 10^{-8} of the solar constant. Fan and Schulte,[11] on the basis of energy emission from the aurora, estimate that nearly 10^8 protons/cm^2/sec enter the auroral zone during a moderately intense aurora. Taking these particles to have a velocity of about 3,000 km/sec (Meinel[12]) at the height of the aurora, we find that they have a kinetic energy of about 10^{-5} of the solar constant. Kiepenheuer,[13] on the basis of "corpuscular E layer" ionization, finds that about 10^{-6} of the solar constant is the energy put into the atmosphere by solar corpuscles. Thus, from these computations, particles also represent only a small fraction of the total solar energy and are absorbed high up in the atmosphere, near the 100-km. level. However, it should be mentioned that Menzel,[14] without giving details, has estimated that as much as 1 solar constant can enter the auroral zone during an intense aurora for a limited period of time. These large uncertainties in the order of magnitude of the energy in the solar particle emission as well as in the amount of irregular solar energy in the ultraviolet will probably be settled when measurements from the earth satellite become available.

Solar Constant.—Of course, variations in direct solar heating may occur near the earth's surface if the "solar constant," as observed by the Smithsonian Institution at the ground, varies. However, the observed solar-constant variations have been questioned; recently Hardie and Giclas[15] compared sunlight reflected from planets with light from nearby stars and found little or no variation in solar emission in visible light. Moreover, the fluctuations which did exist were not correlated with the Smithsonian measurements, so that here again uncertainty exists about the magnitude and even the reality of the measurements of solar-radiation variations.[16]

Variable Sun—Indirect Effects.—Certainly the circulation of the ionosphere and possibly of the ozonosphere will be affected by the absorption of abnormal X-ray and ultraviolet solar radiation in those regions of the atmosphere. If, however, the energies which directly heat the small mass of the upper atmosphere seem too small to affect the lower atmosphere significantly through dynamic effects, solar-weather enthusiasts can point to indirect effects. Numerous ones can be cited, but only one will be singled out for comment here.

Long before the beginning of the twentieth century, it had already been suggested that cirrus clouds and cloudy days were more frequent during the period of sunspot maximum. Numerous additional papers have since purported to show a similar effect. For example, Barber[17] observed the zenith sky with photocells and suggested that the zenith-sky brightness increases with geomagnetic activity; he further suggests that this may indicate water-vapor condensation around solar particles. By contrast, however, Dubois,[18] using Danjon's technique for measuring earthshine on the moon, has found that the earthshine is larger during sunspot minimum than during sunspot maximum. He actually shows the brightness of the dark side of the moon to be nearly 3 times as bright in 1954 (sunspot minimum) as in 1946 (near sunspot maximum). This, in turn, seems to imply a very large change in the albedo of a considerable portion of the earth. If this is due in part to changes in cloudiness, we note that Dubois's results are opposite in sign to many of the earlier findings. Dubois's measurements would be affected most by events in tropical latitudes, whereas the other results are usually (but not always) for more

northerly latitudes. The possibility exists, therefore, that the results may still be consistent with each other. On the other hand, there may be a basic inconsistency here, which is possibly due to the fact that many of these investigations were not statistically significant.

Some experimental support may, however, be found for a cloudiness effect. Crane and Halpern[19] have shown that condensation will occur readily in that portion of a cloud chamber which has been irradiated with weak ultraviolet light. More recently Baum has found at Ohio University (Williams[20]) that a cloud chamber irradiated with ultraviolet light of $\lambda < 4000$ A produces nuclei on which water droplets condense. Unfortunately, these nuclei may be too small to be of importance in the humidity fields which exist in the atmosphere. Yet the fact that ultraviolet emissions can produce nuclei, coupled with Pettit's findings that energy at $\lambda = 3200$ A may be enhanced under disturbed solar conditions, does give some basis to the idea that cirrus clouds might be more frequent under suitable conditions during periods of abnormal solar emission.[21] If such a cloud-producing effect of abnormal solar emissions is substantiated, then the major absorption of the solar energy in the visible and near infrared could be effected, so that considerable change may result in the radiative balance of the atmosphere. Such an upset of the radiation regime, if long continued, could have pronounced effects on the circulation of the lower atmosphere. But obviously much additional work remains before any unambiguous results can be found.

Nonvariable Sun.—Suppose, now, that even the indirect, *variable-sun* effects should prove unimportant for the lower atmosphere. The radiation regime of our atmosphere would nevertheless still undergo large variations. Even if the sun did not vary one iota, large changes in the atmospheric circulation would occur over short periods of time because of the inherent instability of the heated, rotating earth-atmosphere system. The concomitant storminess will have a marked effect on the state of the atmospheric cloudiness, which in turn will have an important effect on the albedo, or reflectivity, of the planet earth. These changes in the albedo cause variations in the solar energy absorbed which are much larger (Danjon,[22] Dubois[18]) than those variations caused by emissions from the disturbed sun as discussed above. Danjon, for example, finds a seasonal variation of albedo in visible light from 0.3 in August to 0.5 in October. If this is correct, the solar energy absorbed varies by at least 25 per cent even on a seasonal basis. These atmospheric-induced variations would, however, not be organized geographically in the same way as the variable solar emissions. The process of cloud formation and dissipation suggests a feedback mechanism by which the atmosphere keeps its own motions regulated to a certain extent. The solar-radiation gradient produces atmospheric motions which set up cloud regimes; these cloud regimes in turn modify the radiation and thus eventually influence the motions themselves.

Clouds.—Since one of the basic controls of the variation of solar radiation in the lower atmosphere is cloudiness, it might be of interest to look more closely at the physics of the interaction between clouds and solar radiation. We might then later see what is required to assess the world-wide effect of clouds on solar radiation absorbed in the atmosphere. Consider clouds of water droplets which are large by comparison with the wave length of sunlight. For such a cloud we may assume (Fritz[23]) that the scattering by the cloud droplets is independent of the wave length of the light; this is borne out by the grayness of thick overcast skies, in marked

contrast to the blueness of the cloudless sky, where spectral scattering is of course very important. Moreover, large spherical water drops have a large forward scatter; nevertheless, as one proceeds into the cloud, the energy gradually becomes more and more diffuse or nearly isotropic. We may think of the diffuse energy as being generated by the scattering influence of the drops on the direct solar beam and on the forward scattered radiation. Let us now divide the cloud into thin layers, each $L/4$ in thickness, where L $(=1/N\pi r^2)$ is the mean free path, N is the number of drops per cubic centimeter, and r is the "radius" of the drops; in a thickness $L/4$ approximately one scatter will occur. Computing the effect of the three-dimensional scatter, we find the generation of diffuse energy as a function of cloud depth. This distribution is shown in Figure 3. The diffuse energy

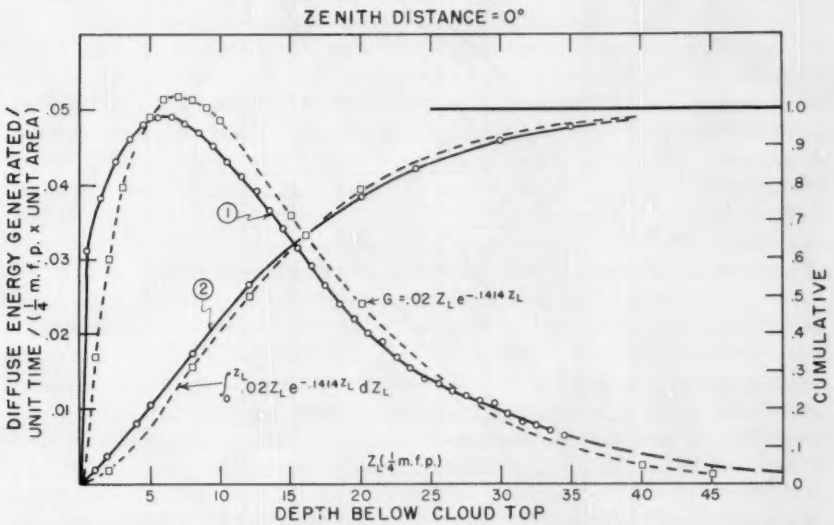


FIG. 3.—The rate of generation of diffuse energy as a function of depth below the cloud top when the solar zenith distance is 0°. Depth is given in units of ¼ mean free path. Solid curves are based on numerical computation; dashed curves are mathematical approximations to the solid curves.

generated per unit volume increases rapidly near the top of the cloud and then decreases gradually. If we neglect absorption in the cloud, the diffuse energy which is generated must diffuse either upward or downward, and a simple diffusion equation enables us to calculate the amount of energy which escapes through the cloud top,[23] as well as the amount which escapes through the bottom (Fritz[24]). Figure 4 shows the albedo, or the fraction of the incident energy which is reflected by the cloud top, when the reflectivity of the underlying earth's surface is 0.1. As we might expect, the albedo depends on the cloud thickness h and the optical properties of the cloud expressed through the mean free path, L. Moreover, especially for thin clouds, the albedo also depends on the sun's zenith distance, Z.

Thus we see that if we want to specify the albedo of an overcast cloud, we should know something about its geometric thickness and about its optical properties. This we rarely know.

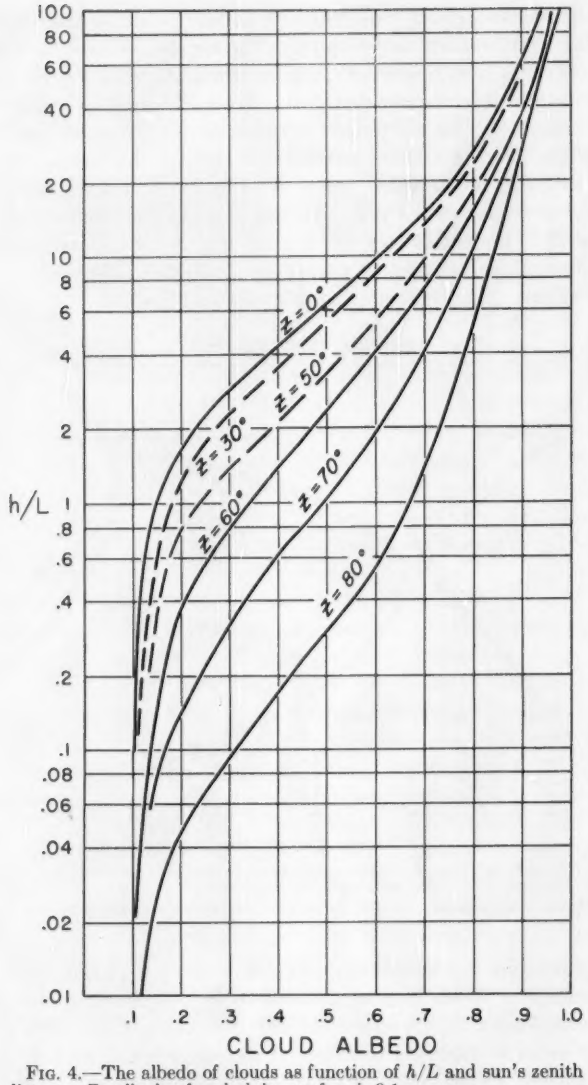


FIG. 4.—The albedo of clouds as function of h/L and sun's zenith distance, Z; albedo of underlying surface is 0.1.

When the sky is only partly cloudy, the reflecting or transmitting properties of clouds become even more difficult to determine on a theoretical basis because of the loss of solar energy through the sides of the clouds. For this problem, we would have to examine empirical data of transmitted energy to the ground as a function of the cloud amount; neglecting absorption in the cloud, which will not usually be the controlling factor, the more energy which is transmitted to the ground the smaller will be the amount reflected by the cloud "top."

Because of the detailed information required, the world-wide distribution of the earth's albedo would be nearly unattainable by observing the clouds from the ground, especially over short periods such as a day or a week. We would not only have to know the cloud types and cloud thicknesses, but we would also have to know the fraction of sky cover over the oceans and over the land. The best way to get the geographical distribution of albedo in the foreseeable future is by means of the earth satellite. But, without waiting for that, we can utilize Danjon's[22] method of comparing the brightness of the dark side of the moon to the brightness of the sunlit side. Since the dark side of the moon is illuminated by sunlight reflected to it by the earth, Danjon's method is a measure of the albedo of that portion of the earth which can be seen from the moon. Through frequent observations from several stations, the method offers a promise of furnishing the longitudinal variation of the earth's albedo. However, the latitudinal variation seems unattainable by this method, because the moon "sees" essentially a lune on the earth which runs more or less from pole to pole. Certain other difficulties are also inherent in the method. For example, near full moon, but when measurements are still possible, the moon will see a portion of earth which is illuminated at grazing angle. Because of specular reflection by water, clouds, and other reflectors on the earth, the earth's albedo may be too large. Further discussion may also be cited (Fritz[25]). But the fact remains that Danjon's method is at present the best, if not the only one, now available for getting an estimate of the earth's albedo from day to day. Perhaps during the International Geophysical Year a program at six to eight observatories will be implemented to furnish the data.

Summary and Conclusion.—We have discussed solar variations and their possible implications for the lower atmosphere. From what is known about variable solar emission, many meteorologists would conclude that the enhanced radiation is too small and is absorbed by too small a mass at the top of the atmosphere to influence the lower atmosphere significantly. But this has not yet been proved. Unknown large variations in the solar spectrum where ozone absorbs (Wulf and Deming) or in particle emissions (Menzel) may perhaps be important. Indirect effects, such as nuclei formation in the presence of enhanced ultraviolet or particle emission, may facilitate cloud formation in the lower atmosphere and change the heat balance of the atmosphere.

But even if the sun were constant, cloud variations would occur which would in turn greatly modify the solar energy absorbed by the earth and its atmosphere. The difficulty of observing the required cloud parameters has been mentioned, and until the earth satellite becomes available, Danjon's method offers the most immediate promise for observing the earth's albedo.

It should, however, be stressed that even after all the radiation parameters are known in great detail, whether it be from the variable sun or from the constant sun by way of cloud variations, the meteorologist still has the formidable task of calculating the influence of the radiation on the atmospheric circulation. With regard to the variable sun, no computations have yet been made. Computations with and without heating should be made with high-speed computers in an attempt to determine the physical plausibility of some of the causes suggested for solar-weather relations. For the normal heating regime of the lower atmosphere with a nonvariable sun, a start has recently been made on computations (Phillips[26]),

although even here the heating regime used in the computational model is a much simpler one than exists in the atmosphere.

[1] B. Duell and G. Duell, *Smithsonian Misc. Collections*, Vol. **110**, No. 8, 1948.

[2] H. Wexler, in *Climatic Change*, ed. H. Shapley (Cambridge, Mass.: Harvard University Press, 1953).

[3] E. D. Farthing, *Bull. Am. Meteorol. Soc.*, **36**, 427, 1955.

[4] F. S. Johnson, *J. Meteorol.*, **11**, 431, 1954.

[5] H. E. Newell, *High Altitude Rocket Research* (New York: Academic Press, Inc., 1953).

[6] E. O. Hulburt, *Naval Research Lab. Rept. 4600*, 1955.

[7] E. Pettit, *Astrophys. J.*, **75**, 185, 1932.

[8] A. P. Mitra, *Ionospheric Research* (Sci. Rept., No. 60 [Pennsylvania State University, 1954]).

[9] O. R. Wulf and L. S. Deming, *Terrestrial Magnetism and Atm. Elec.*, **43**, 283, 1938.

[10] C. De Jager, *Hemel en Dampkring*, **53**, 153, 1955.

[11] C. Y. Fan and D. H. Schulte, *Astrophys. J.*, **120**, 563, 1954.

[12] A. B. Meinel, in *Geophys. Research Paper No. 30*, ed. N. C. Gerson, T. J. Keneshea, and R. J. Donaldson, Jr. (University of Western Ontario and Geophysics Research Directorate, 1954).

[13] K. O. Kiepenheuer, in *The Sun*, ed. G. P. Kuiper (Chicago: University of Chicago Press, 1953).

[14] D. H. Menzel, in *Climatic Change*, ed. H. Shapley (Cambridge, Mass.: Harvard University Press, 1953).

[15] R. H. Hardie and H. L. Giclas, *Astrophys. J.*, **122**, 460, 1955.

[16] It should be emphasized that variations in solar emission in visible light *do* of course exist; visible variations such as sunspots and flares attest to that. But whether or not the variations in emission from the entire sun are too small to be measured quantitatively at the ground with presently available techniques is the question at issue.

[17] D. R. Barber, *J. Atm. and Terrestrial Phys.*, **7**, 170, 1955.

[18] J. Dubois, *L'Astronomie*, **69**, 242, 1955.

[19] H. R. Crane and J. Halpern, *Phys. Rev.*, **56**, 232, 1939.

[20] D. Williams, *Final Rept. Contr. AF* **19(604)-516** (Ohio State University Research Foundation, 1955).

[21] However, Atkinson (*Geofis. pur. e appl.*, **31**, 54, 1955) presents evidence that the experimental nuclei may have been caused by contamination with the rubber in the apparatus.

[22] A. Danjon, *Ann. Observatoire Strasbourg*, **3**, 139, 1936.

[23] S. Fritz, *J. Meteorol.*, **11**, 291, 1954.

[24] S. Fritz, *J. Opt. Soc. Amer.*, **45**, 820, 1955.

[25] S. Fritz, in *Compendium of Meteorology*, ed. T. F. Malone (Boston, Mass.: American Meteorological Society, 1951).

[26] N. A. Phillips, *The General Circulation* (Mimeo Rept. Contr. N-6-ori-139(01) [Princeton, N.J.: Institute for Advanced Study, 1955]).

*ARE CHLOROPLASTS SEMICONDUCTORS?**

By William Arnold and Helen K. Sherwood

BIOLOGY DIVISION, OAK RIDGE NATIONAL LABORATORY, OAK RIDGE, TENNESSEE

Communicated by C. B. van Niel, October 13, 1956

INTRODUCTION

Chlorophyll in photosynthetic organisms is found in small solid bodies known as "grana." In 1941 Szent-Gyorgyi[1] suggested that the protein in the grana might act as a semiconductor, the energy used in photosynthesis being carried by electrons in the conduction band. Eight years later Katz[2] elaborated the scheme and suggested that the electrons may move through a two-dimensional chlorophyll "crystal." In 1955 Bassham and Calvin[3] discussed the same idea and pointed out that an advantage of the semiconductor arrangement would be the physical separation of the points at which reducing agent and oxidizing agent are formed. A second advantage would be that it gives a simple answer to the question of how energy is transferred for long distances in the grana.[4]

The present paper describes two new types of experiments made with dried chloroplasts, the results of which may be described in terms of semiconduction. First, dried chloroplasts give glow curves (light emission on heating) very much like those made with crystals. Second, the resistance of the chloroplasts varies with the temperature in the manner of a semiconductor.

MATERIALS AND METHODS

Fresh leaves of tobacco, spinach, beets, or turnips were washed and dried, and the large mid-ribs were removed. This material (50–100 gm.) was ground in a Waring Blendor with 75–125 cc. of cold 0.1 *M* phosphate buffer (pH 7.0) for 30 seconds. The mixture was filtered through several layers of cheesecloth and centrifuged for $1^1/_2$ minutes at 146 × *G* and the supernatant then centrifuged for 7 minutes at 910 × *G*. The chloroplasts were washed twice with cold distilled water and suspended in a small amount of water. This suspension was then painted onto disks of aluminum, copper, or stainless steel (5 cm. in diameter) and dried with a stream of air. The disks were kept in a desiccator over Drierite until used.

For measurement of the glow curves, the disks were clamped to the top of an aluminum rod 5 cm. in diameter and 15 cm. long that contained a 500-watt heater. Connecting the heater to a Variac permitted the temperature of the rod to be raised at any rate desired. Temperature was measured by a thermocouple located in the aluminum a few millimeters below the disk. The rod could be cooled by dipping the lower half into water, dry-ice mixture, or liquid nitrogen. The upper end of the rod was fastened in a light-tight housing containing a photomultiplier. In the light path between the disk and the photomultiplier there were a shutter, a space for colored glass filters, and a glass cell containing a 1-cm. thickness of flowing water. The flowing water prevented temperature radiation from the disk and rod from warming the photocathode and changing the dark current.

Both the photomultipliers used in this work (RCA No. 6217 and Du Mont K1292) gave the same curves; both tubes were used at dry-ice temperature. The anode current was allowed to flow to ground through 10^9 ohms. A vibrating reed was

used to measure the voltage developed across the resistance. For the curves given in this paper, 1 mv. of signal is approximately 10^6 light quanta per second emitted into 2π solid angle by the disks.

For the resistance measurements, the same disk and aluminum rod assembly was used. A piece of brass (19 × 19 × 6.5 mm.) was placed on the surface of the dried chloroplast film, and the resistance between the brass block and the metal disk measured with a "Hi-Meg" bridge. This instrument is essentially a Wheatstone bridge, one arm of which is a very high Victoreen resistor. The unbalance is detected with a vacuum-tube electrometer.

RESULTS

Glow Curves.—Figure 1 gives the intensity of the light emitted by one of the dried chloroplast samples as a function of temperature, as the temperature of the disk is raised at the rate of 14° C. per minute. The sample had been illuminated at room temperature before the experiment but had not been heated before. If the disk is cooled in the dark and reheated, no light is emitted. If, however, the disk is cooled and reilluminated, light is emitted upon heating. The second glow curve is much smaller than that given in Figure 1. Each time the experiment is repeated the glow curve becomes lower and assumes a different shape. The change from one glow curve to the next becomes less with each repetition, until after four or five heatings the change is very slight. Samples stabilized in this way were used for Figures 2 and 3.

Figure 2 is a chart of the glow curves obtained after different periods of illumination by a 500-watt projection lamp placed 30 cm. from the disk. A Pyrex cell with 3 cm. of water was used to remove the infrared. In each experiment, measurement of the glow curve was started within 1 or 2 minutes after the illumination ceased.

The Du Mont K1292 photomultiplier is sensitive to light out to about 13,000 A in the infrared. The bottom curve in Figure 2, made with a smoked copper disk in place of the sample, shows black-body radiation at the higher temperatures. The black-body curve is the same on heating or cooling, but the glow curve gives light only on heating.

Figure 3 shows the glow curves obtained after illuminating the sample for 5 minutes with the same light system used for Figure 2. During irradiation the temperature of the sample never varied more than 3° C. from the values given on the curves. Measurement of the glow curves was started immediately after the illumination.

Chloroplasts from the different species gave similar glow curves. The kind of metal used for the disk made no difference.

Resistance Measurements.—Figure 4 gives the resistance of the film of dried chloroplasts as a function of the temperature. Curve *A* was obtained on heating the sample for the first time. The disk had been illuminated at room temperature before the experiment. Curve *B* was made as the sample was cooled to room temperature by dipping the lower half of the rod into cold water. The cooling rate was variable under these conditions. The experiment was exactly repeated, except that the sample was not illuminated, giving curves *C* and *D*.

The sample was illuminated for 10 minutes, to obtain the curves in Figure 5.

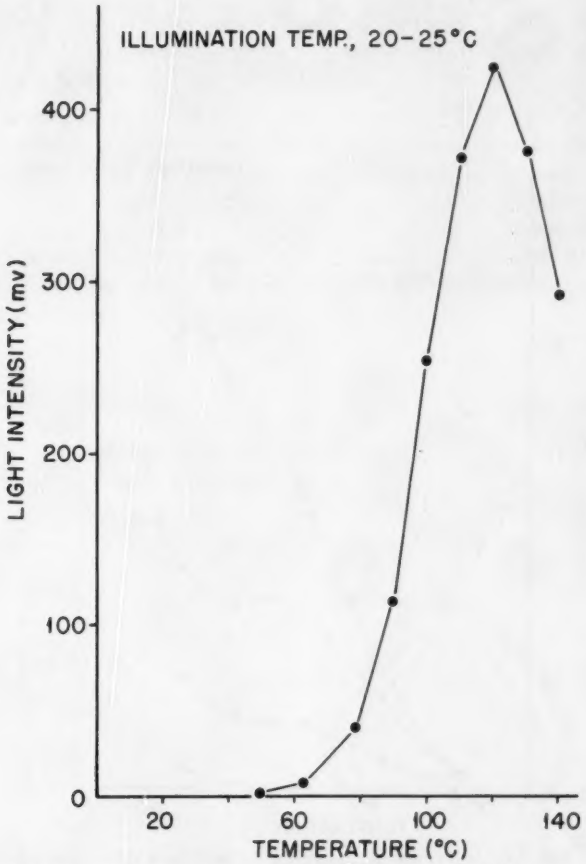


FIG. 1.—Light intensity expressed as a function of the temperature. Integrated light emission up to 140° C. is approximately 8×10^{10} light quanta. (Cf. Fig. 5 in J. T. Randall and M. H. F. Wilkins, *Proc. Roy. Soc. London*, *A*, **18s**, 366, 1945.)

Curve *A* was made after illumination. Curve *B* was obtained as the sample was cooled to room temperature.

DISCUSSION

The experiments described as glow curves show clearly that when the samples of dried chloroplasts were illuminated, a part of the light energy was stored. The stored energy could later be released by heating. Since heating causes the emission, the involvement of an activation energy is implied. A simple experiment shows not a single activation energy but rather a distribution of activation energies. A sample of dried chloroplasts, after being illuminated, was allowed to stand at room temperature in the dark for several hours before the glow curve was made. The intensity of the light emitted at 30°–40° C. was greatly reduced, whereas the light

emitted at 100°–110° C. was not changed from the value it would have had if the glow curve had been made immediately after the illumination. Had only a single activation energy been involved, the ordinates would have been reduced by the same factor at all temperatures.

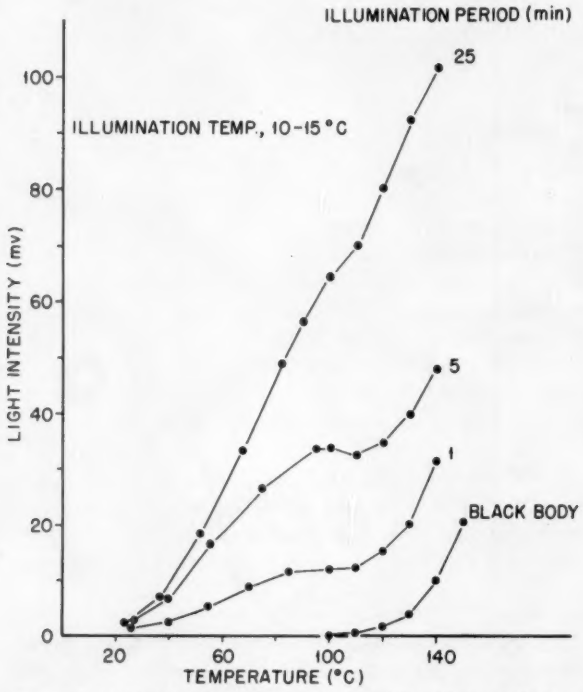


FIG. 2.—Light emission for a sample of "stabilized" chloroplasts as a function of the temperature (rising at 14° C. per minute). The curves show that the amount of light energy stored is not proportional to the time of illumination. Saturation is beginning to show.

An energy storage in crystals analogous to that in the chloroplasts has been known since 1602 from the work of the alchemist Vincenzo Cascariolo.[5] Any of a large number of inorganic crystals, after irradiation with ultraviolet light or X-rays, will emit visible light on being heated. Figure 6 is a schematic representation of the energy levels in the crystal. The absorption of a quantum of ultraviolet light transfers an electron from the filled band into the empty conduction band. As it travels through the conduction band, the electron can be trapped, at certain points in the crystal, by giving up a part of its energy. Once the electron is trapped, it remains at that point in the crystal until a thermal fluctuation supplies the energy needed for it to reach the conduction band. Again, in the conduction band, the electron can be retrapped or it can emit light by falling into a vacant hole in the filled band. The light emission takes place at only certain points in the crystal known as "luminescence centers."

A theory of glow curves has been given by Randall and Wilkins,[6] the equations they derive completely describe the results obtained with chloroplasts.

Experiments made with Corning filter No. 2404 between the 500-watt lamp and the sample showed that the chloroplasts stored almost as much energy as when the whole spectrum was used. Since the filter transmitted only light of wave-lengths longer than 6300 A, and the water cell only those shorter than 14,500 A, the light effective in storing the energy must have a wave-length between these limits. Glow curves made with the No. 2404 filter between the sample and the RCA No. 6217 photomultiplier (long-wave-length limit of sensitivity about 8000 A) show that the wave-length of the emitted light is between 6300–8000 A. These two experiments leave little doubt that it is chlorophyll that absorbs and emits the light in the glow curves.

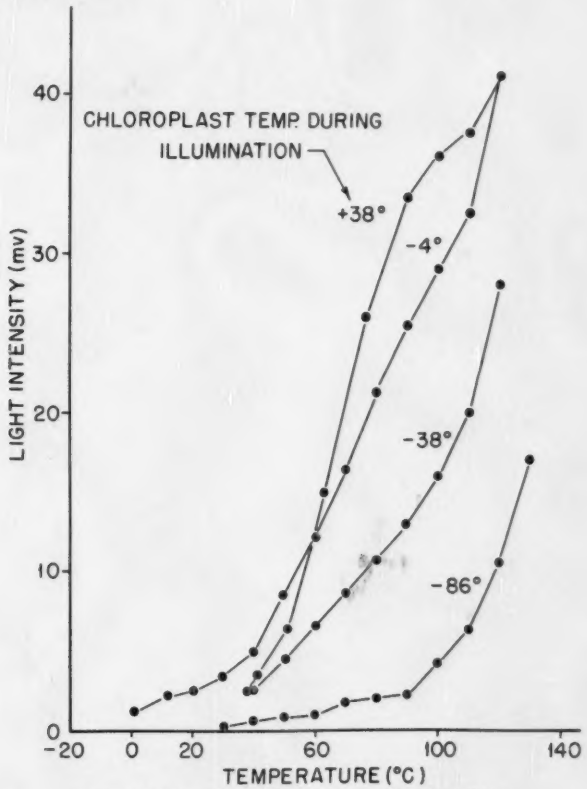


FIG. 3.—Glow curves for "stabilized" chloroplasts.

Figure 3 shows that as the temperature of the sample at the time of illumination is lowered, the amount of energy stored is decreased. After illumination at the temperature of liquid nitrogen, no glow at all is found on heating. At first sight this experiment seems to indicate that since there is a heat of activation for the storage of the light, there must be at least one chemical reaction in the process.

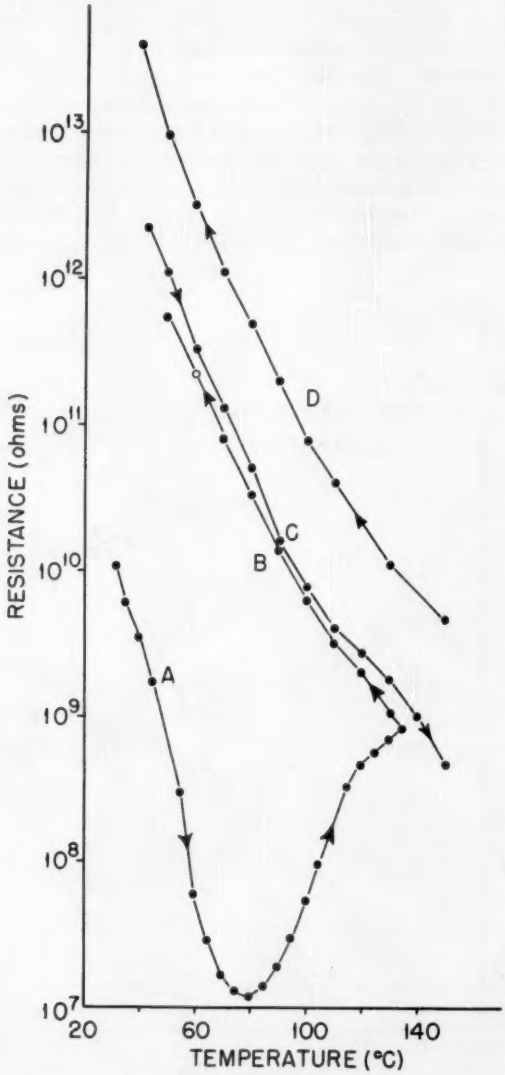


FIG. 4.—Resistance as a function of temperature. Curve *A* was made as temperature of the sample was raised at the rate of 14° C. per minute. Curve *B* was made as the sample was cooled. Curve *C* was made as the sample was heated. Curve *D* was made on cooling the sample down to room temperature, and the points have been raised by one decade in order to avoid congestion on the graph.

However, in solid-state physics, many instances are known where the trapping of an electron involves collision with a phonon, so that, as temperature is lowered, the efficiency of trapping falls. Experiments of the kind represented in Figure 3 cannot help us to decide whether the storage of energy is a chemical reaction or the trapping of electrons.

It now seems certain that the same mechanism that stores energy in the glow curve is also responsible for the delayed light (luminescence) of green plants already

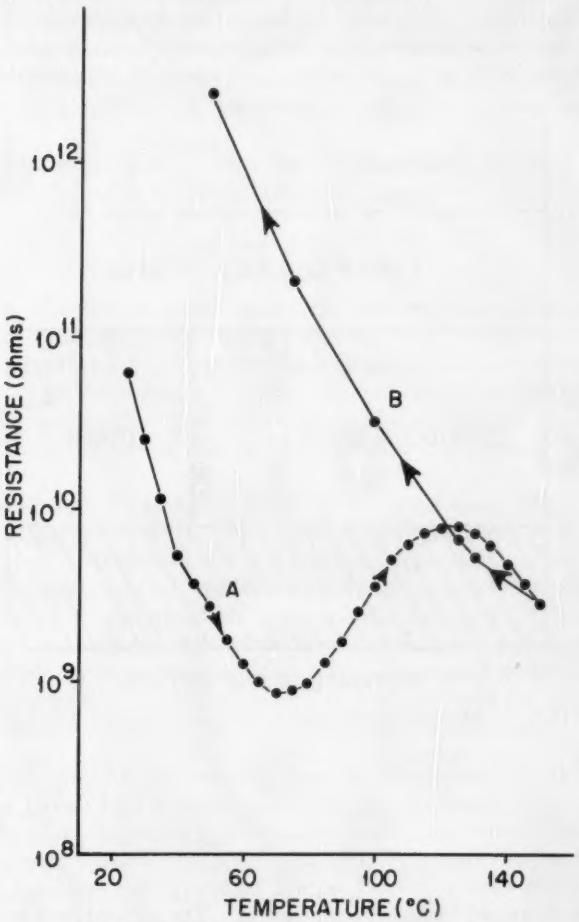


FIG. 5.—Resistance as a function of temperature.

described.[7] When examined with the high-speed phosphoroscope,[8] the dried chloroplast films show the same light emission at short times as do *Chlorella* and leaves. Furthermore, *Chlorella* suspensions and leaf plugs also give glow curves when they are heated after being illuminated. When *Chlorella* and leaves are heated, steam condenses on the optical parts and interferes with the light measure-

ment. The technical problem of making glow curves under pressure to eliminate the fogging has not yet been solved. But enough experiments have been made to say that the glow curves are much the same as with the dried material, except that the peak of intensity seems to be at a lower temperature. The idea that there are a number of light-emitting elements with different activation energies nicely explains the decay curves found for the delayed light.[7, 8]

Crystals that store energy by electron traps show, during the time that the exciting light is on, an increased electrical conductivity. The phenomenon is known as "photoconductivity." If it could be shown that dried chloroplasts are photoconducting, the idea of electron traps being responsible for the glow curves and luminescence would be more believable. A number of experiments have been

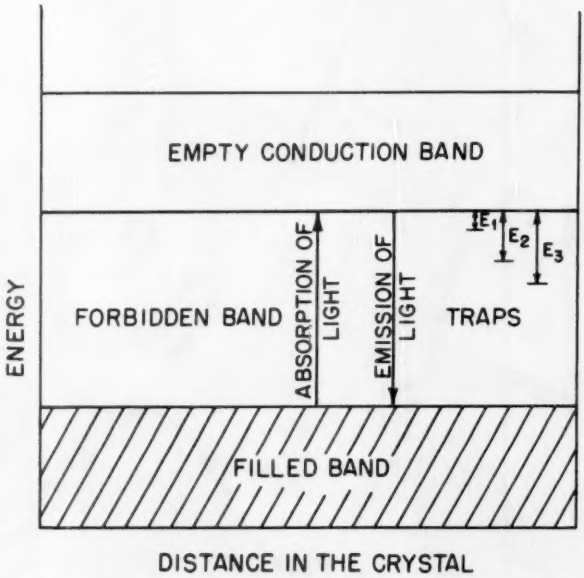


FIG. 6.—A schematic representation of the energy levels in a crystal.

made, but so far we have no unequivocal evidence that red light can increase the conductivity. The experiments do show that infrared light caused a momentary increase in the conductivity. Two explanations of the effect can be given. First, it is well known that infrared light can untrap electrons (transfer them from the trap to the conduction band). Second, as will be shown, the resistance of the dried chloroplasts is decreased by higher temperature. The infrared light would raise the temperature of the sample.

It has been found in Dr. Calvin's laboratory in Berkeley that dried films of chlorophyll are photoconducting (private communication). Since chlorophyll has a high concentration in the grana, it might be argued that the grana must be photoconductive. It should be remembered that the grana are small, discrete bodies imbedded in the stroma of the chloroplasts. Electrical connections made to the

dried material must measure predominantly the electrical properties of the continuous phase.

Since the experiments designed to show that grana are photoconducting were inconclusive, Mr. J. B. Davidson, of this laboratory, suggested that the resistance of the dried chloroplast material be measured under conditions similar to those used for the glow curves.

Figure 4 gives the resistance as a function of the temperature as the temperature of the sample was raised at the rate of 14° C. per minute. The large spike at 80° C., curve A, is a sort of glow curve in conduction. In these dried samples it would seem that this increased conduction must be caused by the freeing of electrons from traps. After the first heating, the spike did not appear again unless the sample had been illuminated as in Figure 5. Experiments with filters showed only blue light to be effective.

The resistance of the samples after the spike has been "annealed out" by the heating follows the equation for an intrinsic semiconductor,[9]

$$\ln R = C + \frac{w}{2kT},$$

where w is the difference in energy between the bottom of the conduction band and the top of the filled band. In Figure 7 curves C and D of Figure 4 have been plotted as $\ln R$ versus $1/T$. From the slope of the line, w has a value of approximately 2.1 ev. This is somewhat smaller than the 4.6 ev. calculated by Gergely and Evans for proteins.[10] That the resistance of dried protein films follows the equation for an intrinsic semiconductor has been known from the work of Baxter[11] and Eley *et al.*[12]

The freeing of electrons shown by the spike in Figures 4 and 5 cannot be closely connected with the glow curves. The spikes in conduction appear at much lower temperature than the peak of intensity in the glow curve, and the glow curve can be made after illumination by red light, whereas blue light is needed for the spike.

The experiments described in this paper do not, unfortunately, give a definite answer to the question "Are chloroplasts semiconductors?" Perhaps experiments on photoconductivity or on the Hall effect may. But, owing to the high electrical resistance shown by the chloroplast samples, such experiments are difficult. However, the similarity between the glow curves made with chloroplasts and those made with crystals, the variation in the resistance of dried chloroplasts as a function of the temperature in the manner of a semiconductor, and the spike in conduction produced by heating all can be understood by assuming that electrons do move in a conduction band in the chloroplast.

If it is established that chloroplasts are semiconductors, then our ideas on the first step in photosynthesis may need some revision.

SUMMARY

1. Glow curves analogous to those made with inorganic crystals can be made with *Chlorella* suspensions, leaves, and dried chloroplasts.

2. On heating, the electrical resistance of dried chloroplasts shows changes that can be interpreted as the freeing of trapped electrons.

3. Although the experiments reported here do not prove that grana act as

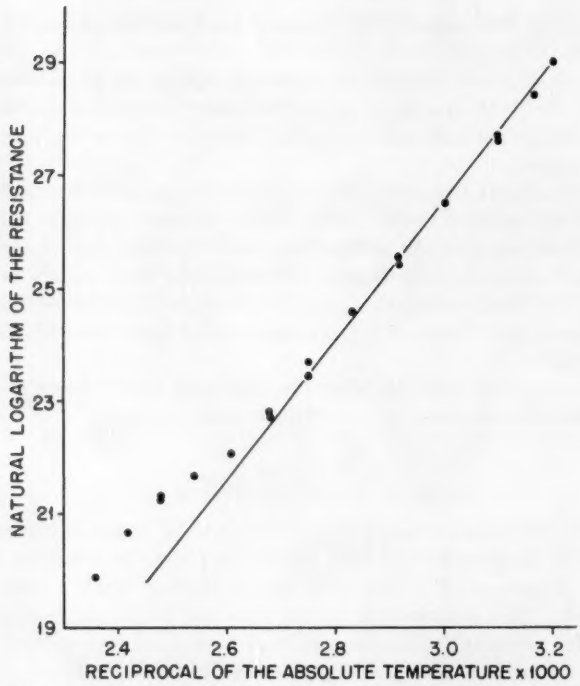


FIG. 7.—The natural logarithm of the resistance as a function of the absolute temperature. The points plotted are those from curves *C* and *D* of Figure 4. If k is taken as 8.6×10^{-5} ev. per degree, the calculated value of w is 2.1 ev.

semiconductors, they do suggest that the first act in photosynthesis may have at least as much in common with solid-state physics as with the chemistry of solutions.

* Work performed under USAEC Contract No. W-7405-eng-26.

[1] A. Szent-Gyorgyi, *Science*, **93**, 609, 1941.

[2] E. Katz, *Photosynthesis in Plants*, (Ames: Iowa State College Press, 1949), chap. xv, p. 291

[3] J. A. Bassham and M. Calvin, USAEC Unclassified Report UCRL-2853, 1955.

[4] W. Arnold and E. S. Meek, *Arch. Biochem. and Biophys.*, **60**, 82, 1956.

[5] *Encyclopaedia Britannica*, article on "Fluorescence and Phosphorescence."

[6] J. T. Randall and M. H. F. Wilkins, *Proc. Roy. Soc. London, A*, **184**, 366, 1945.

[7] B. L. Strehler and W. Arnold, *J. Gen. Physiol.*, **34**, 809, 1951.

[8] W. Arnold, in *Photosynthesis:* Proceedings of the Second Conference (Gatlinburg, Tenn. 1955), ed. H. Gaffron, C. S. French, A. Brown, and N. E. Tolbert (New York: Interscience Publishers, 1957) (in press).

[9] C. Kittel, *Introduction to Solid State Physics* (New York: John Wiley & Sons, Inc., 1953), p. 273.

[10] M. G. Evans and J. Gergely, *Biochim. et biophys. acta*, **3**, 188, 1949.

[11] S. Baxter, *Trans. Faraday Soc.*, **39**, 207, 1943.

[12] D. D. Eley, G. D. Parfitt, M. J. Perry, and D. H. Taysum, *Trans. Faraday Soc.*, **49**, 79, 1953.

STUDIES ON THE IMMUNOLOGICAL OVERLAP AMONG CERTAIN ARTHROPOD-BORNE VIRUSES

*II. THE ROLE OF SEROLOGIC RELATIONSHIPS IN EXPERIMENTAL VACCINATION PROCEDURES**

By Winston H. Price

DEPARTMENTS OF EPIDEMIOLOGY AND BIOCHEMISTRY, SCHOOL OF HYGIENE AND PUBLIC HEALTH, JOHNS HOPKINS UNIVERSITY, BALTIMORE, MARYLAND

Communicated by K. F. Maxcy, September 20, 1956

In vitro neutralization tests, as well as hemagglutination-inhibition and complement-fixation antibody studies have indicated that certain arthropod-borne viruses are immunologically related.[1–3] Four such viruses that appear to be rather closely related are West Nile (WN), St. Louis encephalitis (St.L.), Murray Valley encephalitis (MV), and Japanese B (JB) encephalitis viruses. It was shown by Smithburn[4] that if monkeys were infected with St.L. virus and then with Ilheus virus, which is also related to the above group, such animals developed not only St.L. neutralizing antibodies and Ihleus neutralizing antibodies but also JB and WN neutralizing antibodies. Southam and Moore[5] showed that if humans were given live WN virus and then live Ihleus virus, they formed JB neutralizing antibodies after the infection with Ihleus virus. These results have been interpreted to mean that the above viruses share an antigen with JB virus. Smithburn[4] suggested that such combinations might be used to produce good antibody responses against certain of these viruses.

Work in this laboratory has been concerned with another approach to the problem. This has involved using a live low-virulent virus combined with an injection of a killed vaccine of a related virulent virus in order to produce a good antibody response against certain of the arthropod-borne viruses against which good vaccines have not as yet been prepared. Experiments were reported in which individuals received an injection of live WN virus, followed by an injection of killed JB virus vaccine. Such individuals produced a rapid neutralizing antibody titer to JB virus.[6] The difficulty with the JB killed vaccine now in use has recently been commented on by Hammon.[7] This investigator has also recently presented evidence of the possibility of using the above scheme in vaccinating against certain related arthropod-borne viruses. He showed that individuals and horses in this country who had St.L. antibodies responded with high titers of JB neutralizing antibodies when given one injection of killed JB vaccine.[7]

The data described in this paper present further evidence for the possibility of using in humans a live WN virus in combination with one injection of killed JB virus to protect against several related virulent viruses of the group B classification.[8] The difficulties of using such a vaccination scheme are also presented.

Materials and Methods.—Two WN strains were used. The 101 strain had undergone 4 mouse brain passages. The other WN strain was a strain which had been passed 75 times in minced chick embryo tissue culture and was kindly supplied by Dr. E. Lennette. Neutralization tests were carried out by incubating equal volumes of undiluted unheated sera and varying concentrations of virus for 1 hour at 37° C. and then inoculating 0.03 ml. of the mixture intracerebrally into six 25- to

30-day-old Swiss mice. The mouse LD_{50} was then calculated according to the method of Reed and Muench.[8] Control mice were always included in each titration. The sera to be tested were collected within a few hours after the blood had clotted and were stored at −70° C. in sealed ampules until used. No sera were stored longer than 1 month before being tested, unless otherwise specified. In a few experiments virus dilutions were made in fresh normal human sera in order that any loss of possible cofactors might be overcome. No differences were noted from the data reported in Tables 1 and 2, where the virus dilutions were made in 10 per cent inactivated rabbit sera.

TABLE 1

EFFECT OF HUMANS OF 101 WEST NILE INFECTION ON ANTIBODIES PRODUCED BY SUBSEQUENT INJECTION OF KILLED JAPANESE B VIRUS VACCINE*

	Given West Nile (Maximum Neutralization Index with Indicated Virus)					JB Vaccine Given after Indicated Interval	Given JB Vaccine after West Nile (Maximum Increase in Neutralization Index with Indicated Virus)				
Patient	WN	St.L.†	MV†	JB	Dengue		WN	St.L.†	MV†	JB	Dengue
1-A	$10^{2.1}$	0‡	0	0	0	5 wks.	$10^{3.2}$	$10^{2.0}$	$10^{2.0}$	$10^{4.1}$	$10^{1.1}$
2-A	$10^{3.4}$	0	0	0	0	5 wks.	$10^{4.2}$	$10^{2.0}$	$10^{2.0}$	$10^{3.8}$	$10^{0.8}$
3-A	$10^{2.8}$	0	0	0	0	5 wks.	$10^{3.9}$	$10^{2.0}$	$10^{2.0}$	$10^{3.1}$	0
4-A	$10^{4.1}$	0	0	0	0	5 wks.	$10^{4.9}$	$10^{2.0}$	$10^{2.0}$	$10^{3.2}$	$10^{1.2}$
5-B	$10^{2.1}$	0	0	0	0	9 mo.	$10^{3.4}$	$10^{2.0}$	$10^{2.0}$	$10^{3.4}$	Not done
6-B	$10^{3.1}$	0	0	0	0	$10^{1}/_{2}$ mo.	$10^{3.8}$	$10^{2.0}$	$10^{2.0}$	$10^{3.3}$	Not done
7-B	$10^{4.4}$	0	0	0	0	10 mo.	$10^{4.9}$	$10^{2.0}$	$10^{2.0}$	$10^{2.8}$	Not done

* Samples were collected weekly after the WN virus infection and 4 and 7 days after the vaccine injections and then weekly for as long as possible.

† These values do not necessarily represent maximal indices, as indicated serum was tested against only $10^{1.7}$ LD_{50} of virus.

‡ 0 = Neutralization index of less than $10^{0.8}$.

TABLE 2

EFFECT IN HUMANS OF TISSUE CULTURE WEST NILE STRAIN ON SUBSEQUENT ANTIBODY PRODUCED BY ONE INJECTION OF KILLED JB VIRUS VACCINE

	Given West Nile* (Maximum Neutralization Index with Indicated Virus)		Given Killed JB Vaccine after West Nile Virus* (Maximum Increase in Neutralization Index with Indicated Virus)	
Patient	WN	JB	WN	JB
1-TC	$10^{1.1}$	0	$10^{2.0}$	$10^{2.8}$
2-TC	$10^{1.7}$	0	$10^{2.4}$	$10^{3.1}$
3-TC	$10^{1.2}$	0	$10^{1.9}$	$10^{3.4}$
4-TC	$10^{1.2}$	0	$10^{1.9}$	$10^{2.1}$

* Collection of sera same as in Table 1. JB vaccine given 5 weeks after WN virus.

All patients injected with live WN virus were cancer patients.

Inoculation of Patients with the 101 WN Strain.—Seven patients received $10^{6.7}$ mouse LD_{50} of the 101 strain intramuscularly. These patients responded very similarly to the group of cancer patients injected with this strain by Southam and Moore[5] in so far as clinical symptoms, viremias, and WN antibodies were concerned, and the results will not be reported here. When these patients were given one injection of killed JB vaccine, all of them responded very promptly with increases in JB virus neutralizing antibody, as shown in Table 1. It will be noted that there does not seem to be any correlation between the maximum height of the WN antibody response and the subsequent rise in JB neutralizing antibodies when inocu-

lated with the killed JB vaccine. Patients 1–5 received their JB vaccine 5 weeks after their injection of WN virus, whereas patients 5–7 received their JB vaccine 9, $10^1/_2$, and 12 months, respectively, after their WN injection. It is of interest that no differences were noted in the rate or height of JB antibody formation between these two groups of patients, although the number of individuals involved is too small to permit any conclusions. In all seven cases however, there was detectable JB antibody by the fourth day, reaching a peak at about 7 days. The JB vaccine used in these experiments was formaldehyde-killed mouse brain vaccine, prepared by the Army Medical Service Graduate School and kindly supplied by Dr. Joseph Smadel.

It will be noted that not only did all seven patients show JB antibody rises when given the killed vaccine, but they all showed St.L. and MV neutralizing antibodies, as well as increases in the WN antibodies. It should be pointed out that JB, MV, and St.L. neutralizing antibodies were all absent after the injection of the WN virus in these seven patients under the conditions of the test.[12]

Three of the four patients neutralized small amounts of type 1 dengue virus. These neutralization tests were repeated three separate times. The fact that the patients' sera did not neutralize these small amounts of dengue virus before their one infection of JB vaccine and in view of the evidence presented below that the antibody responses caused by the JB vaccine were not nonspecific, it is felt that the neutralization of the dengue virus by these sera represents a true production of neutralizing antibody to type 1 dengue virus. It is of considerable interest that dengue virus is classified in the same group as WN virus and JB virus.[3]

The antibody response produced by the JB vaccine was not of a nonspecific nature, since there were no increases in the patients' streptolysin O antibodies, in their neutralizing antibodies to influenza A′ and influenza B, or in their polio virus neutralizing antibodies. Furthermore, none of the sera before or after the JB vaccine injection neutralized 30 mouse LD_{50} of Western equine encephalitis. This virus is a member of the A^3 group and is not related to the B^3 group viruses immunologically.

Fifteen control patients who received no WN virus were selected, who were as comparable to the above seven patients as possible. These fifteen patients received their first JB vaccine injection at the same time as patients 1–5 in Table 1 received their WN injections. Five weeks later all twenty patients received one injection of the same JB vaccine used to give the fifteen patients their first injection, the vaccine coming from the same bottle. Of the fifteen patients receiving these two injections of JB vaccine, only 20 per cent showed detectable JB antibodies, and, of these 20 per cent, only 11 per cent showed JB antibodies of comparable titer to those patients receiving WN virus and one injection of JB vaccine.

Injection of Four Patients with the Lennette-Koprowski WN Strain.—One of the main concerns in using the above vaccination procedure is that WN virus, while of low virulence for man, occasionally causes encephalitis. It is therefore possible that if the above procedure were used to immunize large numbers of individuals, too many of the inoculated persons would develop encephalitis to make the above vaccine practical. Experiments in this laboratory[9] have shown that all WN strains so far tested fall into one of three virulent types, based on the reactions they produce in several laboratory animals when inoculated by different routes. The lowest

virulent strain that we have found was that obtained by Lennette and Koprowski on tissue-culture passage, the characteristics of which have been reported in the above-mentioned paper.[9] Four individuals were given intramuscularly $10^{5.8}$ mouse LD_{50} of this virus, which was grown in minced chick embryo as recommended by Koprowski and Lennette.[11] On the basis of clinical and laboratory tests this strain is less virulent in humans than the 101 strain. It caused only 2 days of viremia in two of the four patients, with the maximum viremia titer not exceeding $10^{1.2}$ mouse LD_{50}. None of the patients complained of any clinical illness associated with the virus, and physical examination and laboratory tests showed no deviation from the normal that could be attributed to the virus. One patient had a temperature of 99.8° F. for one day. Groups of 10 *Aedes aegypti* and *Culex pipiens* mosquitoes, which were fed on the four patients daily for 7 days after the WN inoculation, did not pick up any virus when tested by the sensitive means described previously.[6] When the four patients were given their injection of killed JB vaccine, all four showed a quick rise in JB neutralizing antibody titer as shown in Table 2, the maximum titer being reached 1 week after the injection of the vaccine. It would appear on the basis of the small number of individuals tested that, although the tissue-culture strain produces a lower WN antibody response than the 101 strain, it can still sensitize the individual so that he will produce JB neutralizing antibodies when given one injection of killed JB vaccine.

Effect of Previous Injection with 17 D Yellow Fever Vaccine and Live Dengue Virus on Antibody Response Produced by Subsequent Killed JB Vaccine Inoculation.—Since yellow fever virus is in group B[3], it was thought of interest to see whether the 17 D yellow fever vaccine could sensitize humans to a subsequent injection of the killed JB vaccine. If the yellow fever vaccine was similar to WN virus in this respect, then a safe live vaccine would already be at hand. Twelve individuals who were similar to the group used for the WN studies were given the 17 D yellow fever vaccine. Eleven of the twelve individuals developed detectable neutralizing antibodies to yellow fever. However, when part of this group were given one injection of killed JB vaccine 4 weeks after their yellow fever injection and the remaining individuals were given their JB vaccine injection 8 weeks after the yellow fever injection, not one of these individuals developed JB neutralizing antibodies. This indicates that yellow fever virus is not immunologically close enough to JB virus to sensitize the individuals to the JB vaccine under these conditions. This experiment also indicates further that patients of the type being given the JB vaccine are not more sensitive to it than are normal patients, since the individuals in the above study were very similar to those used in the WN study.

Four patients were given $10^{3.0}$ mouse LD_{50} of live type 1 dengue virus (mouse passage No. 6) intramuscularly. Three of these patients developed a rash and several days of fever and headache. All formed good neutralizing titers to dengue virus. When these four patients were given one shot of killed JB vaccine 6 weeks after their dengue inoculation, not one developed JB neutralizing antibodies.[13] It should be emphasized that the JB vaccine used in the negative yellow fever and dengue tests was used at the same time in positive WN experiments.

17 D Yellow Fever Vaccine Plus West Nile Virus.—An experiment was carried out in which seven patients given 17 D yellow fever vaccine were given $10^{7.1}$, mouse LD_{50} of West Nile strain 101 virus four months after the yellow fever vaccine. From

Table III it can be readily seen that these individuals showed a very broad spectrum of neutralizing antibodies against members of the B group. No serum of any patient neutralized approximately 50 LD_{50} of Western equine encephalitis virus (WEE) before or after being given WN virus. WEE is not related to the B group but is in the A group.[3] Preliminary results we have obtained, which will be described in a later paper, suggest that immunization against the group B viruses could be obtained by using the 17 D vaccine plus an attenuated strain of JB virus, or 17 D vaccine plus an attenuated strain of West Nile virus and could be reinforced with one injection of killed JB virus vaccine.

TABLE 3

Neutralizing Antibodies in Individuals Given 17 D Yellow Fever Vaccine Followed by West Nile Virus*

	Maximum Neutralization Index to Indicated Virus after WN Infection*						
Patient	WN	JB	RSS	Dengue 1	Dengue 2	MV	St. L
A. R.	$10^{5.9}$	$10^{2.4}$	$10^{0.8}$	$10^{1.1}$	$10^{0.8}$	$10^{1.8}$	$10^{1.9}$
J. S.	$10^{4.9}$	$10^{1.9}$	0	$10^{0.8}$	0	$10^{1.6}$	$10^{1.4}$
P. H.	$10^{6.1}$	$10^{3.0}$	$10^{1.5}$	$10^{1.2}$	$10^{0.7}$	$10^{2.4}$	$10^{2.6}$
T. E.	$10^{6.1}$	$10^{2.5}$	$10^{1.4}$	$10^{1.2}$	$10^{0.8}$	$10^{1.9}$	$10^{2.1}$
H. G.	$10^{6.3}$	$10^{2.6}$	$10^{1.2}$	$10^{0.7}$	0	$10^{1.4}$	$10^{0.8}$
L. M.	$10^{4.9}$	$10^{2.1}$	$10^{1.1}$	$10^{0.7}$	0	$10^{1.6}$	$10^{1.3}$
H. H.	$10^{5.8}$	$10^{2.8}$	$10^{0.9}$	$10^{1.3}$	$10^{0.8}$	$10^{2.2}$	$10^{1.9}$

* No neutralizing antibodies were found to any of above indicated viruses after 17 D vaccine. The West Nile virus was given 4 months after 17 D vaccine. Patients bled weekly after WN virus. All neutralization tests were carried out by the subcutaneous test in suckling mice except those for dengue virus which were carried out by the intracerebral test. RSS is Russian Spring Summer encephalitis virus. All patients showed neutralizing antibodies to yellow fever virus after being given the 17 D vaccine. Data not shown above indicate that WN infection increased the neutralizing antibody titer to yellow fever virus. 0 values equal a neutralization index of less than $10^{0.7}$.

Discussion.—Individuals given either the live 101 strain or the Lennette-Koprowski strain of WN virus form high titers of neutralizing antibodies to JB when given one injection of killed JB vaccine. The Lennette-Koprowski strain, on the basis of tests in a few individuals, appears to be much less virulent than the 101 strain. Of particular interest was the fact that, although this strain produced much lower WN neutralizing antibodies than the 101 strain, the individuals developed just as high JB neutralizing antibodies when given the killed JB vaccine. The patients who received the 101 WN virus formed antibodies after their one injection of killed JB vaccine that also neutralized St.L. and MV encephalitis viruses. Three out of four of these patients also neutralized small amounts of type 1 dengue virus after receiving their one injection of killed JB vaccine. All patients also showed an increase in WN antibodies after the JB vaccine. This is important, since studies in this laboratory have shown that animals immunized to WN are immune to peripheral challenge with live JB virus and probably MV and St.L. viruses.[9] In this connection, it should be pointed out as emphasized previously[6] that the importance of immunological overlapping between viruses may be overlooked or underestimated unless means of infecting hosts are used that occur in nature unless correct laboratory models are employed.

The JB antibody response produced by the one injection of JB vaccine is detectable 4 days after the injection and reaches a peak about the seventh day. Only 11 per cent of fifteen similar patients who were given two injections of the killed JB vaccine and no WN virus developed neutralizing antibodies comparable to the patients who had one injection of WN virus and one injection of killed JB vaccine.

Indeed, only 20 per cent of these individuals receiving no WN virus formed any detectable JB neutralizing antibodies. The above results are interpreted as being due to the immunological overlapping that occurs between the viruses which share common antigens.

Individuals given either 17 D yellow fever vaccine or type 1 dengue virus did not develop JB neutralizing antibodies when given one injection of the killed JB vaccine. Although both yellow fever virus and dengue virus are immunologically related to the WN-JB viruses, it would appear that this relationship is not extensive enough to sensitize the individuals to the killed JB vaccine under the experimental conditions.

Many problems still remain to be solved. Of paramount importance is the question of how safe the West Nile strain is for humans and how long all the antibodies last. Preliminary data indicate that the JB antibodies induced by the one injection of killed JB vaccine disappear much sooner than the WN antibodies produced due to the infection with live WN virus. It would be of interest to know what happens when such an individual is infected by JB virus, since it is possible that the sensitization that has taken place would cause a quick JB antibody response and protect the individual completely or at least make the infection much milder. Unfortunately, no laboratory animals have as yet been found which would be suitable to test many of the questions that one would like answered. From the results at hand it would appear that under the experimental conditions the Lennette-Koprowski strain of WN virus is incapable of infecting mosquitoes. This would eliminate one possible source of trouble, although it is always possible that there might be a mosquito species in the area where the person was given this WN strain which could be infected. Much more work is necessary on this point, along with work on the best time interval to give the various injections. It is also possible that another virus besides JB should be used in the killed vaccine. Experiments along these lines, combining live WN virus and live and killed dengue vaccine, are in progress. The final objective is to obtain the combination that is safe for human use and which gives the greatest protection against the greatest number of viruses for the longest period of time.

Preliminary experiments indicate that the low virulent-strain of WN can be grown in chick embryos and still retain its low virulence for man. Other data, to be reported in a later paper, also indicate that the JB virus vaccine can also be easily prepared in chick embryos so that it will induce JB neutralizing antibodies when given to humans inoculated previously with the WN virus. Thus the possibility that the vaccinated person would develop encephalomyelitis due to the repeated injection of nerve tissue would be eliminated. Such is not the case with the JB vaccine now in use, which is prepared from mouse brain.

It should be emphasized that the two West Nile strains used in the work reported in this paper would not be suitable for practical application in immunizing humans for a variety of reasons. The data simply indicate that the approach outlined in the beginning of this paper might be feasible for obtaining a satisfactory immunization against certain of the group B viruses. Much more work is needed to develop methods of attenuating and controlling the virulence of West Nile virus along with studies on the homogeneity of such West Nile strains to obtain strains suitable for human vaccination. It should be emphasized that individuals given

certain strains of West Nile virus also routinely develop antibodies to other group B viruses including Japanese B virus[11] so that even without being given an injection of a killed vaccine of a related virulent virus of the B group it is possible that the West Nile infection would reduce the severity of the infection due to certain other viruses in the B group. It should also be pointed out that in this connection an attenuated strain of Japanese B virus would be an excellent one for immunizing humans against certain of the B group viruses, since we have found that infection with Japanese B virus results in higher neutralizing antibodies to other members of the B group than does infection with West Nile virus.[11] The heterologous antibodies that result from the use of either one of the above two strains under the above conditions appear to persist for far longer periods than those induced by the use of the killed virus vaccine procedure as described in this paper, although the use of the killed vaccine greatly increases the neutralizing titers. It should also be noted that even if one assumes low avidity of the heterologous antibodies arising from such infections, these antibodies are capable not only of greatly reducing the viremias of related group B viruses as measured by the titration of the blood in mice, but are also capable of preventing mosquitoes from becoming infected with these viruses.[6]

* This work was carried out under a contract with the Commission on Viral Infections, Armed Forces Epidemiological Board, Department of the Army, Office of the Surgeon General, Washington, D.C. Part of the work under contract with Camp Detrick, Contract No. DA-18-064-404-CML-67.

[1] K. C. Smithburn, *J. Immunol.*, **44**, 25, 1942.

[2] A. B. Sabin, *Federation Proc.*, **8**, 410, 1949.

[3] J. Casals and L. V. Brown, *J. Exptl. Med.*, **99**, 429, 1954.

[4] K. C. Smithburn, *J. Immunol.*, **72**, 376, 1954.

[5] C. M. Southam and A. E. Moore, *J. Immunol.*, **72**, 446, 1954.

[6] W. H. Price, *Public Health Repts.*, **71**, 125, 1956.

[7] W. McD. Hammon, G. E. Sather, E. Lennette, and W. C. Reeves, *Proc. Soc. Exptl. Biol. Med.*, **91**, 517, 1956.

[8] L. J. Reed and H. Muench, *Am. J. Hyg.*, **27**, 493, 1938.

[9] J. Parks, M. Talmadge, and W. H. Price (in preparation).

[10] H. Koprowski and E. Lennette, *J. Exptl. Med.*, **84**, 181, 1946.

[11] W. H. Price, unpublished experiments.

[12] Of twelve individuals given West Nile virus, five described in reference 6 and the seven described in this paper, only two showed JB neutralizing antibody titers as measured by the intracerebral neutralization test before they received their killed vaccine. Dr. Max Theiler, on the basis of experiments he had carried out, suggested that if the subcutaneous neutralization test in baby mice was used, a higher percentage would be found to contain JB neutralizing antibodies after WN infections. We have confirmed his observations.

[13] Dr. A. B. Sabin and Dr. B. Sweet have independently observed similar results in eight individuals given type 1 dengue virus and then one injection of killed JB virus vaccine (personal communication).

THE ORGANIZATION AND DUPLICATION OF CHROMOSOMES AS REVEALED BY AUTORADIOGRAPHIC STUDIES USING TRITIUM-LABELED THYMIDINE

By J. Herbert Taylor,* Philip S. Woods, and Walter L. Hughes

DEPARTMENT OF BOTANY, COLUMBIA UNIVERSITY; BIOLOGY DEPARTMENT AND MEDICAL DEPARTMENT, BROOKHAVEN NATIONAL LABORATORY

Communicated by Franz Schrader, October 26, 1956

Information on the macromolecular organization of chromosomes and their mode of duplication has been difficult to obtain in spite of numerous attempts. One point of attack, long recognized but until recently unattainable, was the selective labeling of some component of the chromosome, the distribution of which could be seen in succeeding cell divisions. Reichard and Estborn[1] demonstrated that N^{15}-labeled thymidine was a precursor of deoxyribonucleic acid (DNA) and that it was not diverted to the synthesis of ribonucleic acid. Recently Friedkin *et al.*[2] and Downing and Schweiger[3] have used C^{14}-labeled thymidine to study DNA synthesis. In chick embryos and *Lactobacillus* there was no appreciable diversion of the tracer to ribonucleic acid. In view of these findings, thymidine appeared to be the intermediate required for the experiment, but the labels so far employed have not been satisfactory for microscopic visualization by autoradiographic means. In order to determine whether an individual chromosome among several in a cell is radioactive, autoradiographs with resolution to chromosomal dimensions must be obtained. Resolution at this level is difficult if not impossible to obtain with most isotopes, since the range of their beta particles is relatively great. Theoretically tritium should provide the highest resolution obtainable, since the beta particles have a maximum energy of only 18 Kev, corresponding to a range of little more than a micron in photographic emulsions. Consequently, identification of this label in particles as small as individual chromosomes should be possible. With this in mind, tritium-labeled thymidine was prepared and used to label chromosomes and to follow their distribution in later divisions by the use of photographic emulsions.

Materials and Methods.—Tritium-labeled thymidine of high specific activity (3×10^3 mc/mM) was prepared by catalytic exchange of tritium from the carboxyl group of acetic acid to a carbon atom in the pyrimidine ring of thymidine (details of the method to be described elsewhere).

Seedlings of *Vicia faba* (English broad bean) were grown in a mineral nutrient solution containing 2–3 μg/ml of the radioactive thymidine. This plant was selected because it has 12 large chromosomes, one pair of which is morphologically distinct, and because the length of the division cycle and the time of DNA synhesisl in the cycle are known.[4] After growth of the seedlings in the isotope solution for the appropriate time, the roots were thoroughly washed with water and the seedlings were transferred to a nonradioactive mineral solution containing colchicine (500 μg/ml) for further growth. At appropriate intervals roots were fixed in ethanol–acetic acid (3:1), hydrolyzed 5 minutes in 1 *N* HCl, stained by the Feulgen reaction, and squashed on microscope slides. Stripping film was applied, and autoradiographs were prepared as described previously.[5]

Experimental Design and Results.—Roots remained in the isotope solution for 8 hours, which is approximately one-third of the division cycle.[4] Since about 8 hours intervene between DNA synthesis in interphase and the next anaphase, few if any nuclei which had incorporated the labeled thymidine should have passed through a division before the roots were transferred to the colchicine solution. In the presence of colchicine, chromosomes contract to the metaphase condition, and the sister chromatids (daughter chromosomes), which ordinarily lie parallel to each other, spread apart. The sister chromatids remain attached at the centromere region for a period of time, but they finally separate completely before transforming into an interphase nucleus. Because colchicine prevents anaphase movement and the formation of daughter cells, but does not prevent chromosomes from duplicating, the number of duplications following exposure to the isotope can be determined for any individual cell by observing the number of chromosomes. Cells without a duplication after transfer to colchicine will have the usual 12 chromosomes at metaphase (c-metaphase), each with the two halves (sister chromatids) spread apart but attached at the centromere. Cells with one intervening duplication will contain 24 chromosomes, and those with two duplications will contain 48 chromosomes.

Two groups of roots were fixed. The first group remained in the colchicine solution 10 hours. The second group remained in the colchicine for 34 hours. In the first group, cells at metaphase had only 12 chromosomes, which indicated that none of these had duplicated more than once during the experiment. The chromosomes in these cells were all labeled, and, furthermore, the two sister chromatids of each chromosome were equally and uniformly labeled (Fig. 1, *a* and *b*). The amount of radioactivity in the chromosomes varied from cell to cell, as would be expected in a nonsynchronized population of cells, but within a given cell the label in different chromosomes was remarkably uniform.

In the second group, cells contained either 12, 24, or 48 chromosomes. Those with 12 chromosomes usually were not labeled, but when labeling occurred, sister chromatids were uniformly labeled as in the first group. In cells with 24 chromosomes, all chromosomes were labeled; however, only one of the two sister chromatids of each was radioactive (Fig. 2, *a* and *b*). Evidently the pool of labeled precursor in the plant had been quickly depleted after the plant was removed from the isotope solution, and these cells with 24 chromosomes had gone through a second duplication in the absence of labeled thymidine.

In the few cells with 48 chromosomes, analysis of all 48 was not possible. However, in several cases where most of the chromosomes were well separated and flattened, approximately one-half of the chromosomes of a complement contained one labeled and one nonlabeled chromatid, while the remainder showed no label in either chromatid. The appearance of cells with 48 chromosomes in a 34-hour period in colchicine also indicates that there was some variation in the predicted 24-hour division cycle.

In cells with 24 and 48 chromosomes a few chromatids were labeled along only a part of their length, but in these cases the sister chromatids were labeled in complementary portions (Fig. 2, *b*, *arrow*). This is the expected situation following sister chromatid exchange and demonstrates that resolution is sufficient to see crossing over in cytological preparations. A careful search of numerous cells with 12 chromosomes failed to yield a decisive case of half-chromatid exchange, which

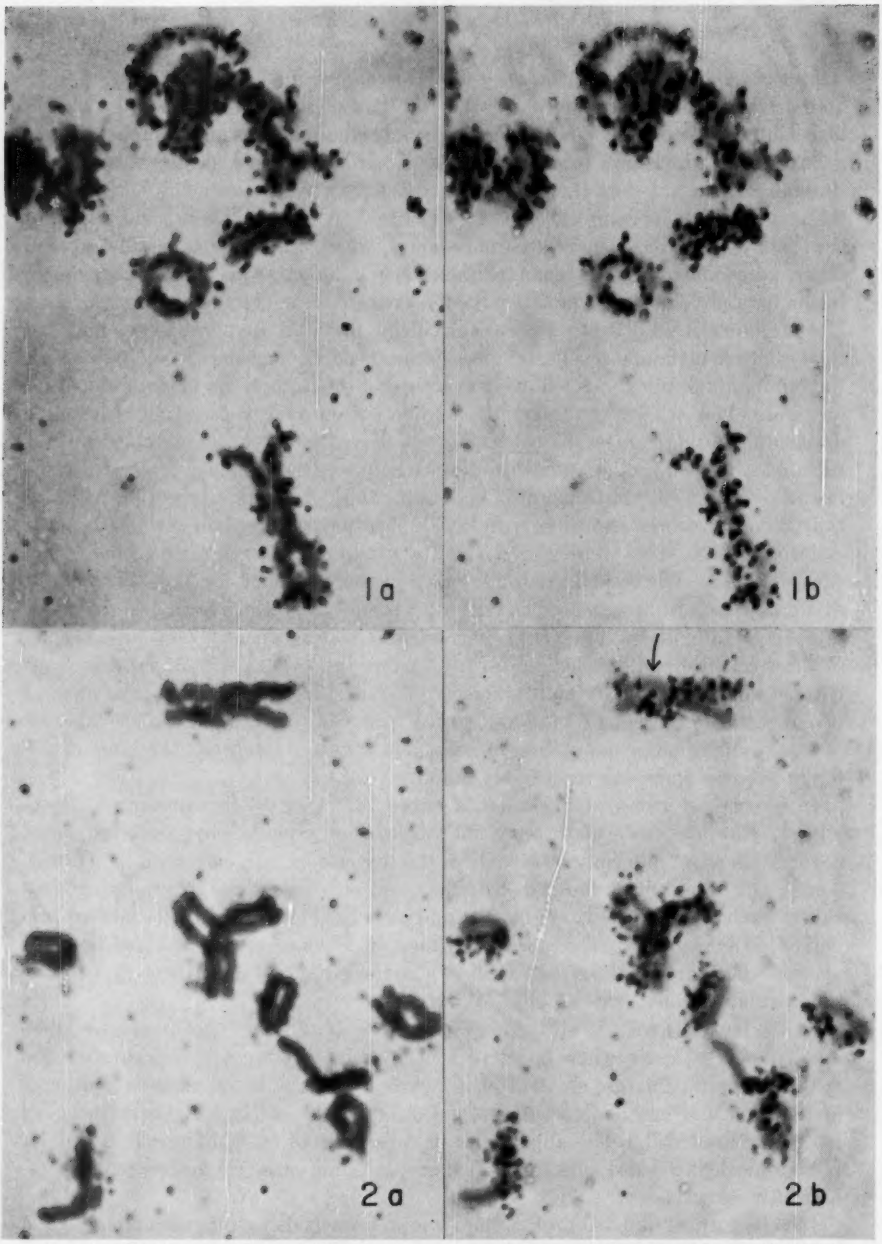


FIG. 1.—Photograph of several chromosomes of a c-metaphase at the first division after labeling occurred; *a*, chromosomes with the chromatids spread apart but still attached at the centromere; *b*, grains in the emulsion above the chromosomes. ×2,200.

FIG. 2.—Photograph of several chromosomes after labeling and one replication in the absence of labeled precursor; *a*, several of the chromosomes from a cell containing 24 chromosomes with chromatids spread but attached at the centromere; *b*, grains in the emulsion above the chromosomes in *a*. ×2,200.

should produce a portion of a chromatid without a label at the first c-metaphase following incorporation of the isotope.

Interpretation and Discussion.—These results indicate (1) that the thymidine built into the DNA of a chromosome is part of a physical entity that remains intact during succeeding replications and nuclear divisions, except for an occasional chromatid exchange; (2) that a chromosome is composed of two such entities probably complementary to each other; and (3) that after replication of each to form a chromosome with four entities, the chromosome divides so that each chromatid (daughter chromosome) regularly receives an "original" and a "new" unit. These conclusions are made clearer by the diagrams in Figure 3. Beginning with two complementary nonlabeled strands in a chromosome, the two strands separate and a complementary labeled strand is produced along each original strand. At the succeeding metaphase each chromatid would appear labeled, although it contains both a labeled and a nonlabeled strand. At a succeeding replication in the absence of labeled precursor, each strand would have a nonlabeled complementary strand produced along its length. At the succeeding metaphase only one chromatid of each chromosome would appear labeled. Following another replication, only one-half of the chromosomes would contain a labeled chromatid, as demonstrated in those cells with 48 chromosomes.

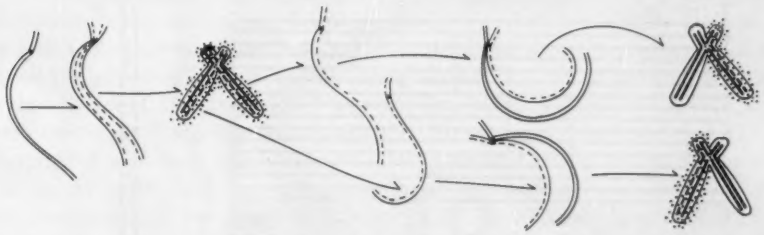

Duplication with labeled thymidine
1st c-metaphase after labeling
Duplication without labeled thymidine
2nd c-metaphase after labeling

FIG. 3.—Diagrammatic representation of proposed organization and mode of replication which would produce the result seen in the autoradiographs. The two units necessary to explain the results are shown, although these were not resolved by microscopic examination. Solid lines represent nonlabeled units, while those in dashed lines are labeled. The dots represent grains in the autoradiographs.

It is immediately apparent that this pattern of replication is analogous to the replicating scheme proposed for DNA by Watson and Crick.[6] We cannot be sure, of course, that separation of the two polynucleotide chains in the double helix is involved, for the chromosome is several orders of magnitude larger than the proposed double helix of DNA.

We know that these large metaphase chromosomes are coiled into at least one helix at the microscopic level and perhaps are twice coiled, a helix within a helix. That the chromosome could be a single supercoiled double helix of DNA is inconceivable when one considers the amount of DNA in a large chromosome. Chromosomes are much more likely to be composed of multistranded units. To explain their duplication as well as their mechanical properties at the microscopic level, they may be visualized as two complementary multistranded ribbons lying flat

upon each other, as shown in Figures 4 and 5. Ribbons of this type with more flexible materials on their outer edges will have a tendency to coil. If the edges contract faster than the central strands when the chromosome begins to shorten, the ribbons fold, one within the other, so as to form a long, trough-shaped cylinder (Fig. 4), and with further contraction they assume the form of a helix. Continued differential contraction would produce a helix within a helix, but the mechanical properties of the model are outside the scope of this discussion.

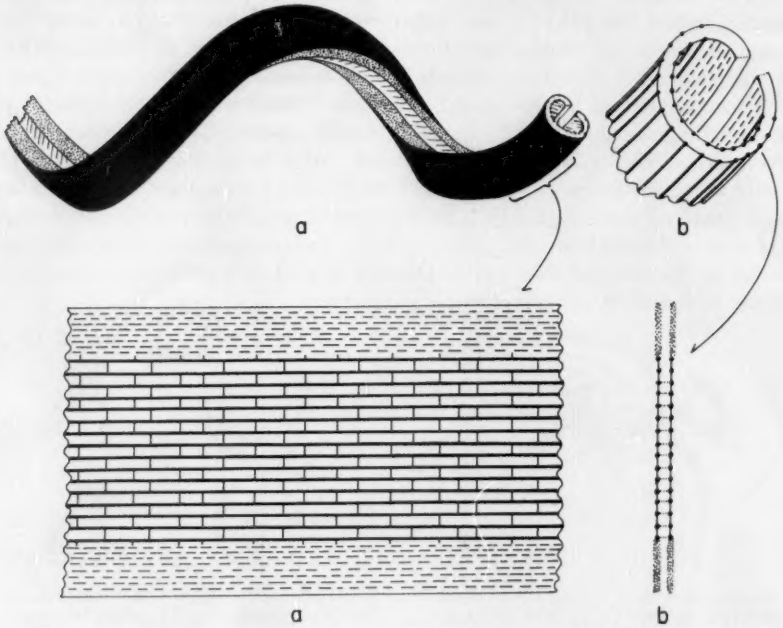


FIG. 4.—Schematic drawing of the proposed ribbon-shaped chromosome with the two multistranded units folded together and coiled; *a*, a single gyre from the coiled chromosome; *b*, detail in cross-section.

FIG. 5.—Diagrammatic sketch of the multistranded units uncoiled and flattened; *a*, short portion in longisection; *b*, cross-section. The number and size of strands shown have no special significance. Although the assumption is made that the strands contain DNA, they do not necessarily correspond to Watson-Crick double helices.

Although the chromosome model is provisional and may require considerable modification and refinement, it has many of the features necessary for duplication and the known stability of genetic materials. The large surface area exposed when the two complementary ribbons are extended would facilitate their rapid duplication. A double-stranded unit with two complementary faces has a high stability,[7] and if the two complementary units are composed of multiple, identical strands cross-bonded, the stability of the larger units should be even greater. Such large units would have a high probability of being transmitted as physical entities. If separation of the complementary faces involves the separation of intertwined double helices of DNA, unwinding presents a problem, but perhaps not an impossible one.[8]

The findings reported here are consistent with those recently reported for the distribution of P^{32}-labeled phage DNA by Levinthal.[9] His data indicated that about 40 per cent of phage DNA is contained in one piece which is divided equally in the formation of two daughter particles, but undergoes no further distribution during the production of about 150 particles that result from the infection of a bacterium. A phage particle would be analogous to a chromosome before duplication, and the first two daughter particles, each of which would be labeled, would be analogous to the two sister chromatids. Since the chromosome is much larger and contains much more DNA than the phage, it is remarkable that they behave in a similar manner in distribution of their DNA during replication.

Our findings are at variance with the report by Mazia and Plaut[10] which was based on the analysis of the anaphase distribution of chromosomes labeled with C^{14}-thymidine. Their data, obtained by the estimation of number of grains over pairs of anaphase or telophase nuclei, indicated a segregation of labeled and nonlabeled units at the first division following the incorporation of the isotope. It is entirely possible that in their experiment more than one division occurred between the time of incorporation and the time the telophase nuclei were analyzed. If this had been the case in the present experiment, unequal distribution of activity in sister chromatids would have been observed in diploid cells.

Summary.—Tritium-labeled thymidine was prepared and used for labeling chromosomes during their duplication. Analysis of autoradiographs showed that both daughter chromosomes resulting from duplication in the presence of labeled thymidine appeared equally and uniformly labeled. After an ensuing duplication in the absence of the labeled DNA precursor, the label appeared in only one of each two chromatids (daughter chromosomes). These findings indicate that DNA is synthesized as a unit which extends throughout the length of the chromosome. The units remain intact through succeeding replications and nuclear divisions, except for occasional chromatid exchanges. Each chromosome is composed of two such units, probably complementary to each other. After each replication the four resulting units separate, so that each daughter chromosome always contains an "original" and a "new" unit. To explain the results, a model with two complementary units and a scheme of replication analogous to the Watson-Crick model of DNA is proposed.

* This work was initiated and the original experiments carried out while the senior author was a research collaborator in the Biology Department, Brookhaven National Laboratory. It has been continued at Columbia University under Contract AT (30-1)-1304 with the Atomic Energy Commission.

[1] P. Reichard and B. Estborn, *J. Biol. Chem.*, **188**, 839, 1951.

[2] M. Freidkin, D. Tilson, and D. Roberts, *J. Biol. Chem.*, **220**, 627, 1956.

[3] M. Downing and B. S. Schweigert, *J. Biol. Chem.*, **220**, 521, 1956.

[4] A. Howard and S. R. Pelc, *Exptl. Cell Research*, **2**, 178, 1951.

[5] J. H. Taylor and R. D. McMaster, *Chromosoma*, **6**, 489, 1954.

[6] J. D. Watson and F. H. C. Crick, *Nature*, **171**, 964, 1953; *Cold Spring Harbor Symposia Quant. Biol.*, **18**, 123, 1953.

[7] H. Kacser, *Science*, **124**, 151, 1956.

[8] N. Ardley, *Nature*, **176**, 465, 1955; M. Delbrück, these PROCEEDINGS, **40**, 783, 1955; G. Gamow, these PROCEEDINGS, **41**, 7. 1955; J. R. Platt, these PROCEEDINGS, **41**, 181, 1955; D. P. Bloch, these PROCEEDINGS, **41**, 1058, 1955; C. Levinthal and H. R. Crane, these PROCEEDINGS, **42**, 436, 1956.

[9] C. Levinthal, these PROCEEDINGS, **42**, 394, 1956.
[10] D. Mazia and W. Plaut, *Biol. Bull.*, **109**, 335, 1955.

A METHOD FOR INCREASING BRAIN SEROTONIN WITHOUT INCURRING SOME OF THE PERIPHERAL EFFECTS OF THE HORMONE

BY D. W. WOOLLEY, E. VAN WINKLE, AND E. SHAW

ROCKEFELLER INSTITUTE FOR MEDICAL RESEARCH, NEW YORK CITY

Communicated November 7, 1956

The object of the present study was to determine whether the serotonin content of the brain could be markedly increased by peripheral medication, but without incurring the peripheral actions of this hormone (contractions of the intestines, increased blood pressure, etc.) at the same time. A further object was to determine whether the increased brain serotonin would then bring about behavioral changes in the animals—in other words, whether the excess hormone would still exert a central effect when some of the receptors in the periphery were blocked. The reasons for this study were the following.

In 1954 Woolley and Shaw[1, 2] presented evidence which indicated that certain mental disorders (e.g., schizophrenia) probably arose from a deficiency of serotonin in the brain. This deficiency was pictured as arising from a metabolic failure of the central nervous system to produce enough of this hormone. The nature of the evidence was such as to implicate serotonin in the etiology of the disease, but it did not unequivocally point to a deficiency rather than an excess of the hormone.[2] One of the purposes of the present investigation was to develop a method which might help distinguish between the alternate possibilities of too much or too little serotonin in the brain as the causative condition of the disorders.

The earlier work had shown[1, 3] that serotonin, although it was formed both in the brain and in the carcass of an animal, did not pass readily between these two parts. Serotonin injected peripherally could not be detected by chemical procedures in the brain, apparently because it did not pass the blood-brain barrier. As a result, there seemed little hope of attempting to raise the serotonin content of the brain by peripheral administration of the hormone. This difficulty was clearly stated in the original paper, and subsequent clinical trials have substantiated it. For example, the peripheral injection of serotonin into patients suffering from schizophrenia has not alleviated the disease. Even if the causative factor in the disorder is a deficiency of serotonin in the brain, the mere administration of the hormone in the periphery could hardly be expected to increase the content of the brain. Direct injection of the hormone into the central nervous system is too dangerous to be contemplated in human subjects. Consequently, a means of changing brain serotonin levels by peripheral medication must be sought if the fundamental hypothesis is to be tested.

A means of doing this has recently been found. Udenfriend *et al.*[4] have found that the normal precursor of serotonin, viz., 5-hydroxytryptophan, when injected peripherally into animals, passes the barrier and is converted into excess serotonin

in the brain. Consequently the peripheral administration of this precursor has been found to increase the serotonin content of the brain.

Nevertheless, a serious difficulty attends the use of this precursor. Not only does it increase the serotonin content of the brain; it also markedly increases the hormone in the peripheral organs. The results of this are, among others, to increase blood pressure by vasoconstriction and to cause strong contractions of the intestines. We have found that the intestinal effects have been particularly noticeable. In animals (dogs, mice) profound diarrhea soon follows administration of 5-hydroxytryptophan. In human subjects Udenfriend[5] has been dissuaded from further trials because of the profound discomfort of these intestinal contractions.

If a way could be found to protect the periphery from the excess serotonin which is formed when 5-hydroxytryptophan is administered, but without also protecting the central nervous system, then the means of testing the original hypothesis with relation to mental disorders would be available. For example, if one had an effective and specific antagonist of serotonin which, however, did not penetrate into the brain, the administration of this antagonist peripherally, followed by 5-hydroxytryptophan also given peripherally, should result in an increase in effective serotonin content of the brain. The peripheral actions of serotonin, however, should be prevented by the specific antagonist which would be confined there. The whole organism, both central nervous system and periphery, would be flooded with excess serotonin from the 5-hydroxytryptophan, but the actions of this excess hormone would be prevented in the peripheral tissues, while the actions in the central nervous system would be unhampered. The selective effect would result from the failure of the antagonist to pass freely into the brain.

The recent discovery of an antimetabolite of serotonin which is highly active against some of the peripheral actions of serotonin, but which seems not to penetrate freely into all parts of the central nervous system, now makes feasible a test of the considerations outlined above. This antimetabolite is 1-benzyl-2,5-dimethylserotonin, the benzyl analogue of serotonin, or BAS.[6, 7] The chief evidence which indicates that it passes into the brain with difficulty is that it does not cause psychiatric disturbances when given to humans. At doses which are able to antagonize several of the peripheral actions of excess serotonin (increase in blood pressure, contractions of the intestines), BAS did not disturb the mind. By contrast, several other antiserotonins have been shown to cause hallucinations or mental depressions.[1, 3]

The object of the present study was to determine whether the peripheral administration of BAS would protect animals from some of the peripheral effects of excess serotonin which arise from injected 5-hydroxytryptophan. The principal peripheral effect to be studied was intestinal contraction, as evidenced by diarrhea. At the same time, the concentration of serotonin in the brain was to be measured to determine whether the hormone was actually being increased in this organ. The central action of this excess cerebral serotonin was to be judged from the behavior of the animals.

Sources of Materials.—BAS or 1-benzyl-2,5-dimethylserotonin hydrochloride was synthesized according to the method previously published by this laboratory.[9] The 5-hydroxytryptophan was the racemic mixture synthesized according to Ek and Witkop[10] and purchased from the California Foundation for Biochemistry.

For injection the BAS was dissolved in water rather than in Ringer's solution because of its much greater solubility in the former. The 5-hydroxytryptophan was dissolved in Ringer's solution.

Methods.—Adult Swiss female mice (25–30 gm. each) were divided into groups of twelve. Each mouse of the BAS groups was injected intraperitoneally with an aqueous solution of BAS, 1 mg. in 0.25 ml. Controls received Ringer's solution. Twenty hours later the injections were repeated. It was very important to pretreat with BAS before the administration of 5-hydroxytryptophan because the BAS was known to exert its maximal antiserotonin effect slowly.

Immediately after the last injection of BAS or saline, 5-hydroxytryptophan (either 2.5 mg. or 15 mg. per mouse) was injected intraperitoneally in 0.5 ml. of Ringer's solution. The animals were placed in individual glass jars and observed very closely for 1 hour. Their behavior was compared to that of controls which had received only saline or only BAS. They were then killed with chloroform, the brains were quickly removed, and analyzed for serotonin by the method of Udenfriend *et al.*[11] This method was scaled down to accommodate the small samples available. This was done by taking larger aliquots of the butanol extract which contained the serotonin, and by use of only 1.2 ml. of acid for extraction from the heptane-butanol mixture. Because serotonin is rather unstable in alkaline solution, it was necessary to minimize the total time of exposure to pH 10 (7 minutes) and to insure that all samples and all standards were exposed for the same time. If this was not done, duplicate analyses frequently did not agree well. With these modifications the method for serotonin determination was found adequate to give good recoveries of added serotonin and yielded values similar to those given by pharmacological methods.

Results.—The data of Table 1 show that the serotonin content of mouse brains was not much changed by prior treatment of the animals with BAS alone. However, the additional administration of 5-hydroxytryptophan brought about a large increase. A dose of 2.5 mg. (100 mg/kg) raised the content appreciably, while 15 mg. caused a 10-fold increase. It was thus clear that mice treated with BAS and 5-hydroxytryptophan underwent a large increase in cerebral serotonin content.

TABLE 1

EFFECTS OF BAS AND 5-HYDROXYTRYPTOPHAN ON DIARRHEA AND CEREBRAL SEROTONIN CONTENT OF MICE

No. of Mice	BAS (Mg/Mouse/day)	5-Hydroxytryptophan (Mg/Mouse)	Incidence of Diarrhea (Per Cent)	Cerebral Serotonin* (μg/Gm)
44	0	0	0	2.3
17	1.0	0	0	2.8
3	0	0.5	100	..
10	0	2.5	100	..
25	1.0	2.5	0	3.4
9	1.0	15.0	0	21.0

* Based on serotonin creatinine sulfate, not on the free base.

When the incidence of diarrhea is taken as a criterion of an effect of serotonin on the intestines, the data of the table clearly show that all of the animals were fully protected from this peripheral effect of the hormone. Prior studies had shown that BAS likewise protected animals from another peripheral action of serotonin, viz., the pressor effects.[7]

The diarrhea in mice receiving only 5-hydroxytryptophan was very severe. It began within a few minutes of the injection and continued for at least 2 hours. The same phenomenon has been noted in dogs. In view of the fact that 0.5 mg. of the compound per mouse caused a maximal response, the ability to protect all the animals which had received 30 times this dose (15 mg. per mouse) was impressive.

The behavioral changes in mice given both BAS and 5-hydroxytryptophan were indicative of profound effects on the central nervous system. This, then, was evidence that the increased cerebral serotonin was in fact still free to act in the central nervous system. Those animals which had received BAS and 2.5 mg. of the hydroxytryptophan showed a very marked quieting or tranquillization. They tended to stand rather close to the ground and not to move at all. They were not asleep, because they would walk when touched, but would soon resume a tranquil state. Control mice which had received the BAS alone without the hydroxytryptophan were noticeably more quiet than normal controls, but they were not affected nearly as much as the ones which had both compounds.

The tranquillity of the mice receiving BAS plus 2.5 mg. of hydroxytryptophan was in marked contrast to the behavior of those getting BAS plus 15 mg. These presented a picture of excitement. They looked rapidly from side to side with great agitation and constantly pranced in one spot and exhibited strong tremors. Their ears became red and their tails stiff. The hind legs were extended behind them. A few walked backward, and, while the behavior was distinguishable from that of mice given lysergic acid diethylamide,[12] the whole picture was somewhat similar to such LSD animals. Most of the mice treated with BAS plus 15 mg. of hydroxytryptophan seemed on the verge of convulsions, and, when they were killed with chloroform, the forward half of the body became very rigid immediately.

It was noteworthy that mice which had received BAS plus hydroxytryptophan did not urinate as normal mice did or salivate as did mice given serotonin intracerebrally.

Behavioral Changes Caused by BAS Injected into the Brain.—In view of the absence of any behavioral change aside from lethargy brought about by peripheral BAS, the effects of intracerebral injection of this compound were tested. When mice were anesthetized with ether and BAS was injected into the lateral ventrical of the brain, pronounced behavioral changes were found. After the anesthesia had been dissipated, the animals began to show principally signs of excitement. About half an hour after the injection, they began to push backward and then to rear up on their haunches and fall over backward. Some of them walked backward in a manner similar to that of mice given lysergic acid diethylamide.[12] Occasionally convulsions were seen. All showed marked reddening of the ears. The behavioral changes appeared much later after the injection than was the case with mice given lysergic acid diethylamide, and the signs became progressively worse. Many animals died after about an hour. These effects were obtained by injection of 50–100 μg. per mouse and were studied in nine mice, all of which exhibited some of the changes. In most of the animals which had received intracerebral BAS, more time was required to awaken from ether anesthesia than with controls given only Ringer's solution.

At 30 μg. per mouse the excitement of the animals did not seem as prominent as at the higher doses. They still pushed backward and reared on their haunches,

but seemed more stunned than excited. This dose represents the amount of BAS which should have reached the brain after peripheral injection of 1 mg., provided that there had been equal distribution to all parts of the body. The behavioral change caused by the intracerebrally injected compound was in sharp contrast to the lack of such change in peripherally injected mice.

Discussion.—These results indicate that when 5-hydroxytryptophan was given to a mouse previously treated with BAS, there was complete protection against some of the effects, notably the contractions of the intestines. On the contrary, the central actions of the hydroxytryptophan were not prevented. Since both peripherally and centrally the hydroxytryptophan appears to act by being converted into serotonin, the present method allows one to increase the cerebral serotonin and to maintain some of its pharmacological effects on the brain and yet at the same time to avoid some of the actions of the excess serotonin in the other organs. It cannot be said whether or not all of the actions of serotonin on peripheral tissues have been circumvented by the BAS. This point must await further study. Nevertheless, one of its most objectionable effects, viz., the diarrhea, was completely eliminated by the antimetabolite.

The evidence that BAS does not enter the central nervous system freely is not entirely conclusive. The chief point in favor of such a view is that, in human beings, BAS does not cause psychiatric difficulties or hallucinations when it is given in doses large enough to prevent the pressor and intestinal actions of peripherally injected serotonin.[13] Many other antimetabolites of serotonin do cause profound psychiatric difficulties.[1, 2, 8] The success of BAS in avoidance of psychiatric effects is presumably due to exclusion of it from some parts of the brain. Nevertheless, this success could also be due to some failure of BAS to combine with the serotonin receptors of the brain, even though it does combine with those in several other tissues. However, the observation that direct injection of BAS into the brains of mice brought about a profound change in behavior would speak in favor of poor penetration from the periphery. Up to the present it has been impossible to decide this question by direct chemical measurement of how much does actually get into the brain, because of failure to recover BAS added to tissues. Until satisfactory recoveries can be achieved, one must be content with the indirect evidence. The results of the present study indicate either that the BAS did not penetrate readily into all parts of the brain or that it was poorly able to antagonize serotonin in some of its functions in the brain. By contrast, one piece of evidence exists to show that BAS does have some sort of action on the central nervous system. In humans and in mice, large doses have a tranquillizing effect.

The marked changes in behavior of BAS-treated mice when the dose of 5-hydroxytryptophan was altered are noteworthy. With 2.5 mg. per mouse the picture was one of deep tranquillization or lethargy. With 15 mg. it was one of great excitement approaching convulsions. Our own earlier studies in mice and those of others with dogs, rabbits, and rats given serotonin by direct installation into the brain have shown that the most prominent effect is to induce lethargy. As the dose is increased to very large amounts (compared to that normally present in the brain), convulsions appear. It is possible (for theory see Shaw and Woolley[7, 14]) that these large doses are acting specifically to block and inactivate the serotonin receptors.

Summary.—When mice were given the specific antimetabolite of serotonin, viz., 1-benzyl-2,5-dimethylserotonin or BAS, followed by 5-hydroxytryptophan, the serotonin content of their brains was markedly increased. This was reflected in behavioral changes and thus was an indication that the excess serotonin in the brain was pharmacologically active. Some of the effects on peripheral organs, however, were prevented. This was true of the diarrhea, due to intestinal contractions, which 5-hydroxytryptophan caused to a marked degree in unprotected animals. The relationship of these results to experimental psychiatry has been indicated.

[1] D. W. Woolley and E. Shaw, These PROCEEDINGS, **40**, 228, 1954.
[2] D. W. Woolley and E. Shaw, *Brit. Med. J.*, **2**, 122, 1954.
[3] E. Shaw and D. W. Woolley, *J. Pharmacol. Exptl. Therap.*, **111**, 43, 1954.
[4] S. Udenfriend, D. F. Bogdanski, and H. Weissbach, *Federation Proc.*, **15**, 493, 1956.
[5] S. Udenfriend, personal communication.
[6] D. W. Woolley and E. N. Shaw, *Science*, **124**, 34, 1956.
[7] E. N. Shaw and D. W. Woolley, *J. Pharmacol. Exptl. Therap.*, **116**, 164, 1956.
[8] D. W. Woolley and E. Shaw, "Symposium on Psychotomimetic Drugs," *Ann. N.Y. Acad. Sci.*, April, 1956.
[9] E. Shaw, *J. Am. Chem. Soc.*, **77**, 4319, 1955.
[10] A. Ek and B. Witkop, *J. Am. Chem. Soc.*, **76**, 5579, 1954.
[11] S. Undenfriend, H. Weissbach, and C. T. Clark, *J. Biol. Chem.*, **215**, 337, 1955.
[12] D. W. Woolley, These PROCEEDINGS, **41**, 338, 1955.
[13] R. W. Wilkins, *New England J. Med.*, **255**, 115, 1956.
[14] E. Shaw and D. W. Woolley, *Science*, **124**, 121, 1956.

SOME FACTORS INFLUENCING THE LONG-WAVE LIMIT OF PHOTOSYNTHESIS

BY ROBERT EMERSON, RUTH CHALMERS, AND CARL CEDERSTRAND

BOTANY DEPARTMENT, UNIVERSITY OF ILLINOIS, URBANA, ILLINOIS

Communicated November 19, 1956

There is evidence from several sources that the quantum efficiency of photosynthesis begins to decline at wave-lengths where light absorption by chlorophyll is still strong. Emerson and Lewis[1] found a sharp drop in efficiency beyond 685 mμ for both *Chlorella* and *Chroöcoccus*. Tanada[2] found a similar decline for the diatom *Navicula minima* at about the same wave length. Action spectra measured by Haxo and Blinks[3] for representatives of the green and brown algae also indicate declining photosynthetic efficiency beyond the red maximum of chlorophyll absorption.

It has seemed plausible to Rabinowitch,[4] Duysens,[5] and others that photosynthesis is brought about entirely by the lowest excited electronic state of chlorophyll *a;* yet this appears to conflict with the evidence of declining photosynthetic activity within the red absorption band of chlorophyll. Absorption anywhere within the red band, followed by vibrational equilibration with the medium, should bring the chlorophyll molecule into the same excited electronic state, with the same (temperature-determined) vibrational distribution (cf. Rabinowitch,

op. cit., pp. 1156–1157). In the living cell, the red absorption maximum is close to 680 mμ, and, although the long-wave side of the band is very steep, absorption probably remains appreciable as far as 700 mμ and perhaps somewhat beyond. The sharp decline in efficiency of photosynthesis beginning at or near 685 mμ therefore requires interpretation. A decline in yield of fluorescence on the long-wave side of an absorption band, known for chlorophyll as well as for many other dye molecules, poses a similar problem (cf. Rabinowitch, *op. cit.*, p. 752, and Stepanov[6]).

We reported,[7] for both *Chlorella pyrenoidosa* and the red alga *Porphyridium cruentum*, that at lower temperatures the decline in yield of photosynthesis begins at longer wave lengths and that in the region of the long-wave decline the yield could be improved by supplementary light of shorter wave lengths. We give here the results of additional experiments with *Chlorella* which confirm these effects of temperature and supplementary light in greater detail.

Our manometric technique for measuring photosynthesis was similar to that described by Emerson and Chalmers,[8] except that we used the single-vessel instead of the two-vessel method. We are concerned here with steady rates, which can be measured by the single-vessel method in either carbonate buffer (about pH 9) or in acid culture medium (about pH 5). In the case of single-vessel measurements in acid culture medium, it is necessary to assume a value for γ (the ratio CO_2/O_2). We have used a value of -1, since it has been well established that γ is close to -1 for steady rates of photosynthesis and respiration of *Chlorella*. The two-vessel measurements of Emerson and Chalmers show that this is specifically the case for the conditions of the experiments to be reported here.

The cell material (*Chlorella pyrenoidosa*) was grown at about 23° C., over a combination of fluorescent and incandescent lamps. The inoculum was 30 μl. of cells in 200 ml. of medium, and the harvest (after 3 days' growth) was about 200 μl. of cells. For the measurements of photosynthesis, about 250 μl. of cells were suspended in 8 ml. of liquid. In this high concentration the cells transmitted no appreciable fraction of the light beam used for measuring quantum yields, out to about 700 mμ.

There is little doubt that, out to this point, absence of transmission implies practically total absorption of the energy of the incident beam by chlorophyll and hence that the number of absorbed quanta equals the number of incident quanta. Visual inspection (the eye remains a sensitive detector nearly to 700 mμ) and tests with an emission-type photocell and amplifier betrayed no appreciable transmission. Beyond 700 mμ, the photocell and amplifier continued to show no evidence of transmission to about 710 mμ, where it became just detectable. In this region, absorption by chlorophyll is probably so small that scattering and reflection begin to contribute appreciably to the quenching of the incident beam, and therefore absence of evidence of transmission becomes an uncertain criterion of totality of absorption. However, our data, presented in this communication, that full efficiency of photosynthesis can be extended to 700 mμ by supplementary light, are strong evidence that we have maintained total absorption to this point. Tests made by Emerson and Lewis[9] with different densities of suspension indicated that total absorption could be maintained to somewhat beyond 700 mμ.

Illumination was from the grating[10] monochromator described by Emerson and Lewis.[11] The light source was a tungsten ribbon filament lamp operated at about

7.5 volts and from 29 to 31 amps. At each wave length the current was adjusted to give equal numbers of incident photons per unit time (about 0.1 μeinstein per minute).

The energy of the incident beam was measured with a bolometer calibrated against radiation standards from the U.S. Bureau of Standards. Appropriate corrections were made for reflection losses at the bolometer window and for differences in optical path to bolometer and to reaction vessel.

The band width was estimated by mounting a Zeiss hand spectroscope at the exit slit of the monochromator and moving the hand spectroscope across the width of the exit slit. The band width did not exceed 10 mμ. This is probably equivalent to the band half-width of 5 mμ specified by Emerson and Lewis[11] for their measurements in the red region. It is more appropriate to specify the full visible width, because the wide slits make the significance of the half-width uncertain.

With dense cell suspensions and low light intensities, respiration exceeds photosynthesis, so that the rate of respiration during light exposures is an important correction term in the calculation of quantum yields. The usual method of estimating this correction (interpolation from respiration measurements made during dark periods before and after light exposures) led to small maxima and minima in the plots of quantum yield against wave length. The positions of these maxima and minima depended on the sequence of exposures to different wave lengths and on the spacing of the dark periods. A new method of estimating the respiration during light exposures eliminated the uncertain maxima and minima, as well as the dependence of yield on the sequence in which the exposures to different wave lengths were made. Instead of interspersing many dark periods among the light exposures, we made only one direct measurement of respiration, in a single dark period at about the middle of the experiment. At this time we determined the difference between the rate of respiration in darkness and the net rate of photosynthesis and respiration at a standard wave length (660 mμ). Other experiments had shown that this difference, D, remained nearly constant during the course of an experiment; hence exposures to a standard wave length (always at the same intensity) could be used as an index of changes in rate of respiration, without repeatedly returning the cells to darkness. The observed net rate at the standard wave length was plotted against time, and the difference D was subtracted from the plot to find the curve for respiration as a function of time. The respiration correction applicable to each measurement of photosynthesis was read directly from this curve.

Figure 1 shows two sets of measurements made in this way. The sequence of observations was first from short wave lengths to long (solid circles and continuous line) and then from long wave lengths to short (open circles and dashed line). The agreement between the two sequences is close, and there is no evidence of systematic differences.

Each curve plotted in Figures 2 and 3 represents an average of two sets of observations like those in Figure 1, first in sequence from short wave lengths to long, then in the opposite sequence.

In experiments with supplementary light no direct measurement of respiration was required, but it was nevertheless necessary to make allowance for changes in respiration. The net metabolic rate during photosynthesis from the supplementary

light varied with changes in respiration. This net rate was measured at intervals, and a smooth curve was drawn through the observations. To find the increment in photosynthesis attributable to the measured energy of the monochromatic beam, we took the difference between the observed rate of supplementary light plus monochromator beam, and the rate for supplementary light alone, read from the smooth curve for the corresponding time.

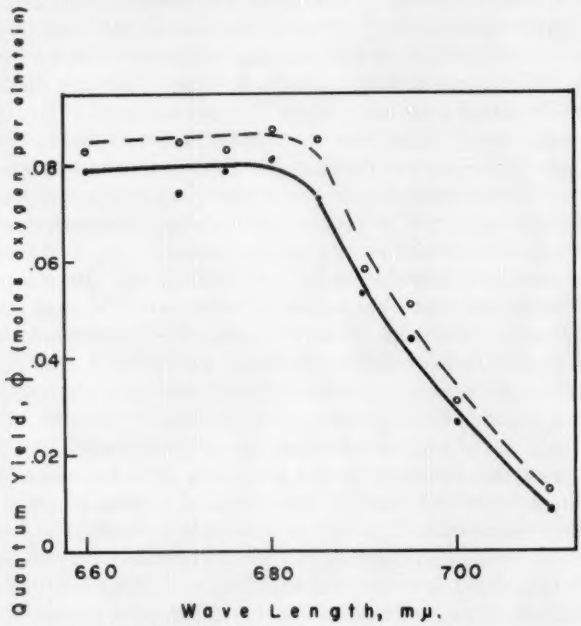
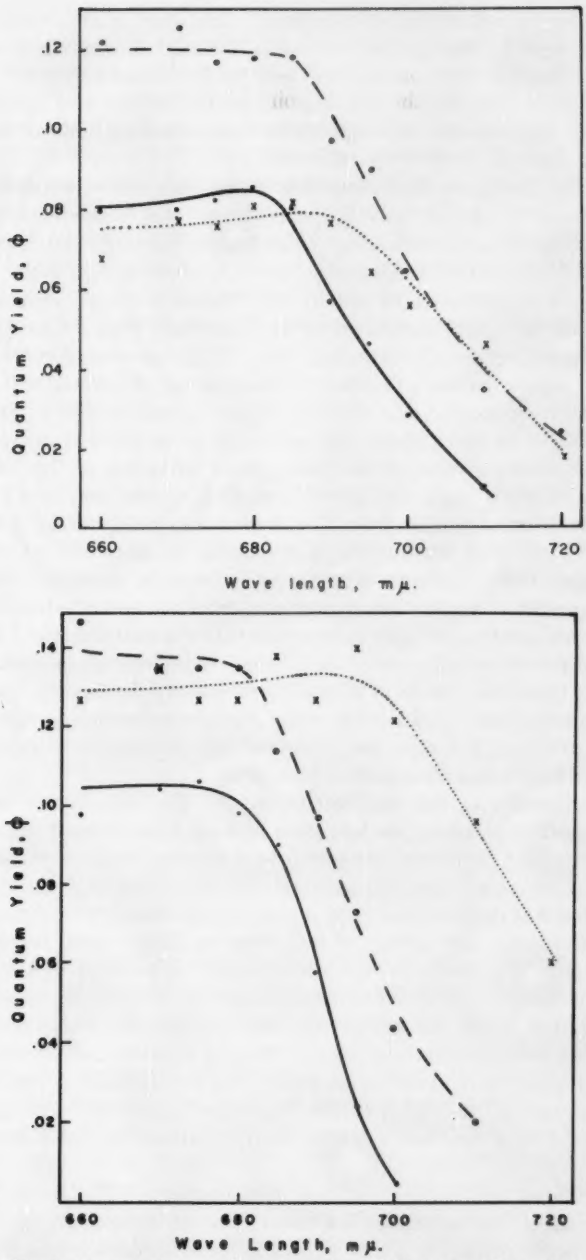


FIG. 1.—A test to show that the method of measuring dependence of yield on wave length gives the same result for opposite sequences of wave lengths. Solid dots and continuous line are for sequence from short wave lengths to long, measured first; measurements in opposite sequence are shown by open circles and dashed line.

The supplementary light was from a mercury-cadmium lamp. The intensity was adjusted to maintain photosynthesis in excess of respiration by an amount roughly equal to the respiration itself. The supplementary light entered the reaction vessel from above, and the monochromator beam from below. The cell suspension was transmissive for the yellow and green lines of the mercury-cadmium lamp, so that all cells in the suspension were continuously illuminated, though at varying intensity, as they circulated. On the other hand, the cells were intermittently illuminated by the monochromator beam; the intensity of this light dropped to zero as it penetrated the suspension.

The effects of temperature and supplementary light on the long-wave limit of photosynthesis are shown in Figure 2 for cells in carbonate buffer and in Figure 3 for cells in culture medium. In both figures the curve for 20° (solid circles and continuous line) bends downward at the shortest wave length, the curve for 5° (open circles and dashed line) at a slightly longer wave length, and the curve at 20° with supplementary light (crosses and dotted line) extends horizontally to the



FIGS. 2-3.—Effect of temperature and supplementary light on quantum yield. Fig. 2, cells in carbonate buffer. Fig. 3, cells in acid culture medium. In each figure the solid dots and continuous line are for 20°, the open circles and dashed line for 5°. The dotted curves and crosses are for measurements at 20° with supplementary light.

longest wave length before turning downward. The wave lengths where the curves decline to one-half their maximum level are in the same sequence and can be more precisely specified than the long-wave limits of the horizontal portions. In Figure 2 the halfway points come at 696, 702, and 711 mμ for 20°, 5°, and 20° with supplementary light, respectively. In Figure 3 the corresponding halfway points are at 690, 695, and 718 mμ.

All the curves in Figure 3 are somewhat higher than the corresponding ones in Figure 2. Rieke[12] also found that the quantum yield was higher for cells suspended in culture medium than for cells in carbonate buffer; hence the differences between Figures 2 and 3 with respect to maximum yield are probably significant. Less easy to explain is the difference, to be seen in both Figures 2 and 3, between maximum yields at 5° and 20°. The dashed curve is substantially higher than the solid line, for both carbonate buffer and culture medium. If further work should confirm that the quantum yield is indeed a function of temperature, this would call for revision of current interpretations. Our method, however, was devised to show the dependence of yield on wave length, perhaps at some sacrifice of comparability in regard to the absolute value of the yield. Our derivation of the correction for respiration from only a single dark period in each experiment may lead to differences in precision at different temperatures, and the value of γ may not be quite the same at 5° and 20°. Since at 20° the rate of respiration is higher and is also subject to greater changes, there is greater opportunity for error in quantum yields through error in estimation of respiration correction. Without special attention to these factors, it would not be justifiable to conclude that the quantum yield is a function of temperature or to attach special significance to differences in maximum yield. On the other hand, the effects of temperature and supplementary light on wave-length dependence were confirmed in many experiments over a period of months, and we regard this as evidence that comparability with respect to dependence of yield on wave length was adequately maintained.

Since the intensity of the supplementary light was sufficient to make photosynthesis exceed respiration, the low yield at long wave lengths without supplementary light might have been characteristic of photosynthesis of these cells below the compensation point. Several authors (for example, Kok,[13] Bassham *et al.*[14]) have suggested that the quantum yield of photosynthesis may be different above and below compensation. The difference they reported (higher yield below compensation than above) is opposite to that which would be required to account for our results, but if there are factors which can change the yield in the neighborhood of the compensation point, the change may not necessarily be always in the same direction. We have therefore tested the effect of intensity, above and below the compensation point, both in the region where yield is not changing with wave length (640–670 mμ) and in the region of the long-wave decline (675–700 mμ). It was necessary to use these wide bands in order to attain intensities well above compensation.

Figure 4 shows the results of these measurements. For both wave-length regions the rate of photosynthesis is a linear function of intensity below, through, and above the compensation point. The slope is lower for the longer-wave band, in agreement with the lower quantum yields shown in Figures 1–3 at longer wave lengths. There is no indication that higher intensities alone can raise the yield for

longer wave lengths. On the other hand, supplementary light of shorter wave lengths can improve the yield for the long-wave band, even at the highest intensity shown in Figure 4, just as it does for the very low intensities of long wave lengths used in Figures 1–3. In Figure 4 the point marked with a square represents the yield for the wave band 675–700 mμ at the same intensity as the solid black dot just below, but supplemented by light of shorter wave lengths. It is raised to the same level as the corresponding point on the line for the 640–670 mμ band.

These measurements showing no departure from linearity above and below compensation ("Kok effect") are in agreement with similar measurements[15] with the red cadmium line (644 mμ). In the course of many measurements under a wide range of conditions, we have never observed significant changes in yield in the region of compensation.

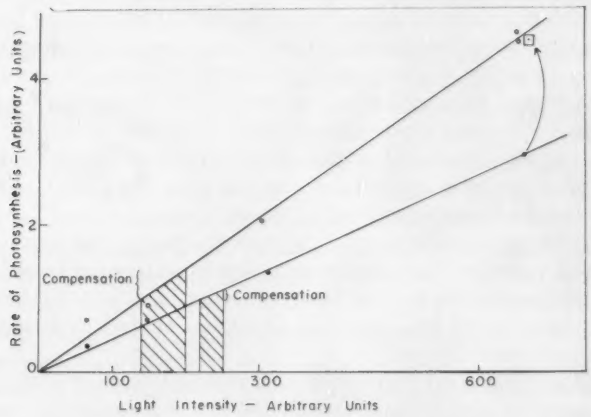


FIG. 4.—Dependence of rate of photosynthesis on intensity. The upper line (*open circles*) is for a band of wave lengths from 640 to 670 mμ. The lower line (*solid dots*) is for 675–700 mμ. The point marked by a square is for the 675–700 band, supplemented by light of shorter wave lengths.

These experiments with higher intensities of light show that the effects of supplementary light on the yield in the far-red region cannot be attributed solely to the higher rate of photosynthesis in the supplementary light. The shorter wave lengths of the supplementary light must be significant. We mention here the results of some preliminary experiments with different wave lengths and intensities of supplementary light, although our work on these factors is incomplete.

By means of colored glass filters, we have isolated groups of lines from the blue, green, or red regions of the spectrum, from the mercury-cadmium source of supplementary light. We found positive effects (i.e., improvement in yield from long-wave red) with all three spectral regions when we used intensities of supplementary light sufficient to give photosynthesis somewhat above compensation of respiration. Intensities too small to give measurable photosynthesis showed no effect on the yield from long-wave red. It seems particularly interesting that effects were obtainable with supplementary red light. This was chiefly the cadmium line at

644 mμ, a wave length still appreciably shorter than the band of long-wave red (680–700 mμ) which, according to Figure 4, gave no improvement in yield at higher intensities. Somewhere between 644 and 680 mμ there must be a limit beyond which the supplementary light is no longer effective.

These exploratory experiments indicate that the effect of supplementary light on the yield from long-wave red differs in important respects from the action of supplementary light described by Warburg *et al.*[16] They reported effects of supplementary light on the yield of photosynthesis from a band of red light centering at about 644 mμ. This is a region of wave lengths appreciably shorter than the long-wave band where we found effects of supplementary light. Our Figure 2 shows no effect for wave lengths shorter than 680 mμ. The dotted line and solid line represent yields at 20° with and without supplementary light. The corresponding curves in Figure 3 show some difference in level for wave lengths shorter than 680 mμ, but the difference is small compared to the effect at longer wave lengths. The measurements in Figures 2 and 3 start at 660 mμ, a wave length longer than the red band at 644 mμ used by Warburg *et al.* However, we have done a number of experiments starting at 640 and 650 mμ and have failed to find significant effects of supplementary light except at wave lengths longer than 680 mμ.

Warburg *et al.* also report that intensities of supplementary light too low to give measurable photosynthesis were effective in increasing the yield for red light and that there was a sharp maximum in effectiveness of supplementary light at 460 mμ. Wave lengths longer than 500 mμ gave very small effects. Our measurements, on the other hand, indicated that intensities of supplementary light sufficient to give measurable photosynthesis were necessary in order to show improvement in the yield from long-wave red light and that all regions of the visible spectrum were effective, at least out to 644 mμ.

Warburg *et al.* report that in red light without supplementary blue, γ for photosynthesis approaches zero, while the supplementary blue restores the normal value of about -1 for γ. Although we have made many measurements of γ for photosynthesis in the red cadmium line (644 mμ), we have never observed γ values approaching zero for steady rates of photosynthesis. In our experience,[8] γ is close to -1, except during periods of transient changes in rate.

We note that Warburg *et al.* measured the effects of supplementary light by the two-vessel manometric method, which can sometimes lead to large errors through inattention to details. We are not prepared to offer a quantitative explanation of how the method might produce the results they describe, but we refer to our earlier analysis[8] of the sensitivity of the two-vessel method to small differences between the vessels. Possibly small differences in light absorption may have contributed to the effects of supplementary light which appear in their results.

Our provisional conclusion is that the effects of supplementary light described by Warburg *et al.* are so different from the effects we have observed that the prospect of a common interpretation is unpromising.

Our results show that for *Chlorella* the effect of temperature on the long-wave limit is not conspicuous. According to Figures 2 and 3, a drop in temperature from 20° down to 5° extends the limit only about 5 mμ. Emerson *et al.*[7] mentioned measurements with the red alga *Porphyridium cruentum* which showed a greater dependence on temperature. The long-wave limit of full efficiency was about 650

mμ at 20° and about 670 mμ at 5°. Besides being farther apart, these limits are at considerably shorter wave lengths than the corresponding limits for *Chlorella* (cf. Figs. 2 and 3).

Haxo and Blinks[3] noticed a similar difference between red and green algae. Instead of measuring quantum yields under conditions of total absorption, they used thin (translucent) pieces of thallus and compared action spectra with absorption spectra. Their figures show that for green and also for brown algae the action spectrum of photosynthesis parallels the absorption spectrum through the red maximum of chlorophyll, while for red algae the action spectrum begins to fall sharply away from the absorption curve at about 650 mμ and drops practically to zero in the red absorption maximum of chlorophyll. They made their measurements at room temperature (probably about 20°), and the wave lengths where their action spectra begin to fall below their absorption spectra correspond reasonably well with the wave lengths where we found declining yields of photosynthesis in our experiments at 20° with *Chlorella* and *Porphyridium*.

Haxo and Blinks concluded that in red algae the light absorbed by chlorophyll is of minor value for photosynthesis and that phycoerythrin plays the major photochemical part. Our measurements at 5° show, however, that in the case of *Porphyridium* the range of full efficiency may extend to 670 mμ, or well into the red absorption band of chlorophyll. If this effect of low temperature proves to hold for red algae in general, there will then be no need to suppose that their chlorophyll is necessarily less efficient than that of green algae. Taking the effect of temperature into consideration, we may say that the analogy between the red and green algae is close. In both cases the range of full efficiency can extend well into the red absorption band of chlorophyll, and the long-wave limit of full efficiency falls appreciably short of the long-wave limit of the red absorption band. It may be possible to find a common interpretation.

A comparison of the limits of absorption and photosynthesis shows that for both red and green algae the yield is low in the region where absorption is attributable to chlorophyll *a* alone. For *Chlorella* the yield remains high to about the wave length where we may expect absorption by chlorophyll *b* to terminate. The long-wave limit of absorption for the *b* component in living cells cannot be directly observed, but it may be estimated from the limit for the *a* component. With proper correction for scattering, measurements with living cells show that at 705 mμ the absorption is close to zero and dropping steeply (Latimer and Rabinowitch[17]). This is the long-wave side of the chlorophyll *a* band. If the distance between the red maxima of the *a* and *b* components is a correct index of the difference between their long-wave margins, then at 690 mμ the absorption due to chlorophyll *b* must be close to zero. Beyond this, absorption must be attributed to chlorophyll *a* alone. Red algae lack the *b* component but are believed to contain small amounts of chlorophyll *d*. Since the concentration of the *d* component is probably so small that it makes a negligible contribution to absorption even in the region of its maximum, we shall overlook it for the moment and consider that the beginning of absorption by chlorophyll *a* alone will coincide with the end of absorption by phycobilins. We may expect this to be variable because red algae contain varying amounts of phycocyanins as well as phycoerythrin, but at least we can say that the zone where chlorophyll *a* alone accounts for practically all the absorption will generally extend

toward shorter wave lengths than for green algae. This is in qualitative correspondence with the observed long-wave limits of full efficiency.

For both red and green algae, supplementary light extends the range of full efficiency into the region of exclusive absorption by chlorophyll *a*, and perhaps to the limit of the *a* absorption band. This is shown for *Chlorella* by the dotted curves in Figures 2 and 3. Marcia Brody is preparing for publication the results of similar measurements with *Porphyridium* which show that, for this alga also, supplementary light extends the long-wave limit to near 700 mμ.

Since the supplementary light must be of shorter wave lengths than those which by themselves give diminished yield, the significance of the supplementary light may be that it adds excitation of other pigments besides chlorophyll *a*. The maintenance of maximum efficiency may require the excitation of some pigment with an absorption band corresponding to an energy level higher than the first excited state of chlorophyll *a*. In the green algae this could be chlorophyll *b*, while in red algae it could be one of the phycobilins. Since the red maximum of chlorophyll *d* is to the long-wave side of chlorophyll *a*, it would not fill the requirement, and this is the primary reason we omitted it from consideration.

The brown algae and diatoms also lack chlorophyll *b* but contain a *c* component in sufficient concentration to contribute appreciably to absorption. According to Tanada,[2] the long-wave limit of full efficiency of the diatom *Navicula minima* is at about 675 mμ (temperature of 10°), or slightly toward shorter wave lengths than the long-wave limit for *Chlorella*, and we may suppose that the limit of chlorophyll *c* absorption is also on the short-wave side of the chlorophyll *b* limit, since this is true of their maxima.

For blue-green algae, which also lack chlorophyll *b*, we would expect the phycocyanins to fill the role we have suggested for chlorophyll *b* in *Chlorella*. Different phycocyanins probably have different long-wave absorption limits. As in the case of red algae, we might anticipate a range of limits for full efficiency. Emerson and Lewis[18] found that for *Chroöcoccus* full efficiency extended to about 680 mμ. Their Figure 1 indicates that absorption by phycocyanin remains appreciable about to this point. Haxo and Blinks, on the other hand, mentioned preliminary experiments with other blue green algae which indicated termination of full efficiency at shorter wave lengths.

We look forward to testing the effect of supplementary light on the long-wave limits for brown and blue-green algae, to find whether they fit the interpretation we have suggested. It is in conflict with the widely accepted view that transfer of excitation energy to chlorophyll *a* from other pigments takes place with practically 100 per cent efficiency (cf. Duysens[5]). It also fails to account for the reduced yield of fluorescence on the long-wave side of the absorption band of chlorophyll *a*. However, if further work should confirm our suggestion that photosynthesis requires excitation of two different pigments, we can hardly expect to find a common interpretation for the long-wave decline in yield of fluorescence and photosynthesis.

We take pleasure in acknowledging support of this research from the National Science Foundation under Grant G1398.

[1] Robert Emerson and Charlton M. Lewis, *J. Gen. Physiol.*, **25**, 579–595, 1942, Fig. 4; *Am. J. Botany*, **30**, 165–178, 1943, Figs. 5 and 6.

[2] T. Tanada, *Am. J. Botany*, **38**, 276–283, 1951, Fig. 4.

[3] F. T. Haxo and L. R. Blinks, *J. Gen. Physiol.*, **33**, 389–422, 1950.

[4] Eugene I. Rabinowitch, *Photosynthesis and Related Processes* (New York: Interscience Publishers, Inc., 1951), Vol. **2**, Part 1.

[5] L. N. M. Duysens, *Transfer of Excitation Energy in Photosynthesis* (Utrecht: Maatschappij v/h Kemink en Zoon N. V., 1952).

[6] B. I. Stepanov, *Izvest. Akad. Nauk S.S.S.R., Ser. Fyz.*, **20**, 493, 1956; B. S. Neporent and N. A. Borisevich, *Izvest. Akad. Nauk S.S.S.R., Ser. Fyz.*, **20**, 476, 1956; *Doklady Akad. Nauk S.S.S.R.*, **69**, 695, 1954.

[7] Robert Emerson, Ruth Chalmers, Carl Cederstrand, and Marcia Brody, *Science*, **123**, 673, 1956.

[8] Robert Emerson and Ruth Chalmers, *Plant Physiol.*, **30**, 504–529, 1955.

[9] Emerson and Lewis, *Am. J. Botany*, **30**, 165–178, 1943, Fig. 6.

[10] We are indebted to the Carnegie Institution of Washington (Mt. Wilson Observatory) for the loan of the grating in the monochromator.

[11] Emerson and Lewis, *Am. J. Botany*, **30**, 165–178, 1943.

[12] F. F. Rieke, "Quantum Efficiencies for Photosynthesis and Photoreduction in Green Plants," chap. 12 of *Photosynthesis in Plants*, ed. James Franck and W. E. Loomis (Ames: Iowa State College Press, 1949), pp. 261–263.

[13] B. Kok, *Biochim. et biophys. acta*, **3**, 625–631, 1949.

[14] J. A. Bassham, K. Shibata, and M. Calvin, *Biochim. et biophys. acta*, **17**, 332–340, 1955.

[15] Robert Emerson and Ruth V. Chalmers, "Research in Photosynthesis," *Gatlinburg Symposium, 1955*, ed. Hans Gaffron and others. New York: Interscience, 1957.

[16] O. Warburg, S. Krippahl, and W. Schroeder, *Z. Naturforsch.*, **10B**, 631–639, 1955.

[17] Paul Latimer and Eugene Rabinowitch, *J. Chem. Phys.*, **24**, 480, 1956.

[18] Emerson and Lewis, *J. Gen. Physiol.*, **25**, 579–595, 1942, Fig. 4.

THE SYMMETRY OF DENDRITIC SNOW CRYSTALS

BY DAN MCLACHLAN, JR.

STANFORD RESEARCH INSTITUTE, MENLO PARK, CALIFORNIA

Communicated by Sterling Hendricks, November 1, 1956

Snow crystals are of interest in two major respects: (*a*) meteorologically and (*b*) artistically. The meteorological interest in snow crystals was greatly enhanced by the work of Langmuir[1, 2] and Schaefer,[3–6] who have done such extensive pioneering work on the seeding of clouds with material having a structure similar to that of ice, with the object of gaining some control over rainfall throughout the country. Their most effective seeding materials are frozen carbon dioxide and silver iodide. From the artistic standpoint the interest in the beauty of snow crystals goes back to the most primitive times, but we know that as early as 1555 Olaus Magnus, Archbishop of Upsala, published a woodcut of a snow crystal in a book on the general subject of "Natural Phenomena." The most authentic early drawings of snow crystals were made by Scoresby[7] in 1820. One of the first to use photography for the study of shapes of snow crystals was Hellmann[8] in 1893. The most extensive work on the photography of snow crystals was carried out by W. A. Bentley, of Jericho, Vermont, who apparently spent a good portion of his life at this task, judging from his publications between 1901 and 1927.[9] In 1931 W. J. Humphreys assisted Bentley in compiling about 2,400 of Bentley's most interesting photographs

into a beautifully arranged book.[10] All recent work on the study of snow crystals is contained in a well-illustrated book by Nakaya.[11] This book lists eighty-seven references and is outstanding for its complete accounts of the experiments of Nakaya and Hanajima in their artificial production of snow crystals to duplicate almost every form produced naturally.

The present discussion deals with the artistic aspects of snow crystals, but it has a scientific interest because it tries to explain the causes of the interesting shapes of snow crystals and, furthermore, because these explanations may be carried over into other fields wherein dendritic growth and branching are evident.

Three outstanding characteristics of dendritic snow crystals are immediately apparent: (*a*) they exhibit a hexagonal pattern; (*b*) the crystals are all different in shape, as though no simple law (such as governs "normal" growth) operated to cause a relationship between the crystals; and (*c*) the branches of the six trees on the dendritic crystals form six almost identical sets. The hexagonal character of snowflakes is in perfect accord with the crystal structure or atomic arrangement of the water atoms in the ice crystal. Bernal and Fowler have worked out the structure of ordinary ice, or ice I,[12] while the structures of ice formed at high pressures have been worked out by Megaw,[13] Burton and Oliver,[14] and McFarlan.[15] The fact that no two snowflakes are alike seems to surprise no one, since experience teaches that the probability of a difference existing between two objects increases with the complexity of detail in the objects. However, the fact that the six trees (see Fig. 6) branching from a common center in a snow crystal are almost identical in spite of their intricacy is a thing that excites the wonder of everyone.

It is the great similarity between the six parts of a dendritic branching snow crystal that this paper is written to explain. The problem has often been expressed by the question, "How does one branch of the crystal know what the other branches are doing during growth?" Such regularity and symmetry are not so uncommon among the plants, especially in the flowers and blossoms, and among the animals, particularly in animals of the sea, where hormones and nerves provide a means of co-ordinating the development and activities of living organisms (Thompson[16]); but for an inanimate organism to duplicate its development in growth in such minute detail is not easily explained. The explanation of the co-ordination in growth among the six branches of a snow crystal is based (in this discussion) upon the theory that it is caused by standing waves of the thermal and acoustical type within the crystal. We will first discuss the nature of the standing waves[17] within bodies of various shapes and then discuss the kinetics of the deposition of molecules of H_2O on the crystal.

The Nature of Standing Waves.—When an elastic substance is disturbed by applying some force momentarily to some portion of it so as to displace that portion from its equilibrium position, that portion of the substance will (as soon as the force is removed) return to its initial position. But, owing to the momentum of that portion, upon its arrival at the initial position, it will "coast" beyond equilibrium. Again in a position of nonequilibrium due to "overshooting," the body again finds itself elastically deformed and returns through the equilibrium position with momentum which carries it to the position of original displacement. This process may be repeated many times, depending upon how rapidly the energy is dissipated by friction, either internal or external. The number of times which this cycle is

repeated per second is called the frequency ν and depends upon the restoring force F and the mass M of the displaced material; for simple systems this may be expressed as

$$\nu = \frac{1}{2\pi}\sqrt{\frac{F}{M}}, \tag{1}$$

where ν is the frequency. At the same time vibrations may be carried to other portions of the body. The number of centimeters per second with which the disturbance is transmitted from one portion of the body to another is called the "velocity" of the wave.

If the body is not infinite in extent but is limited to some defined shape, the waves will be reflected at the boundaries, and the reflected waves will mix with other oncoming waves. There are certain frequencies of vibration whose values ν are such that there is a sort of coherence between all reflected waves and the oncoming waves. For these frequencies, which are called "characteristic frequencies," so-called "standing waves" result. The standing waves appear not to be moving. Their apparent stationary condition results from the fact that the oncoming and reflected waves have equal amplitude and equal frequencies and their velocities are equal but in opposite directions. The frequencies of standing waves depend upon the initial conditions of displacement and upon the shape and dimensions of the body. These conditions are called "boundary conditions." From the boundary conditions the characteristic frequencies may be solved for, provided that the boundary conditions are sufficiently simple.[18, 19] The shapes for which solutions may be obtained are (*a*) linear systems, parallelograms, and parallelepipeds, to which one-dimensional, two-dimensional, and three-dimensional Fourier series apply, respectively; (*b*) circular systems and cylindrical systems, to which Bessel functions apply; and (*c*) spherical systems, to which Legendre polynomials apply.

It may be assumed that a long, unbranched spine of a dendritic crystal of snow complies very nearly to a one-dimensional system. When in transverse vibration, the displacement $\phi(x, t)$ of a point at distance x along its length at any instant of time t can be expressed by the Fourier series

$$\phi(x, t) = \sum_m A_m \cos\frac{2\pi m x}{a}\cos 2\pi\nu_m t, \tag{2}$$

where a is the length of the spine, m is an integer, and A_m is the amplitude of the mth wave. In this respect the spine is analogous to a vibrating string free at both ends. As is known to all musicians, the relative values of the amplitudes A_m for each m value (or harmonic) is influenced (1) by the position x_p at which the string is plucked, struck, or bowed and (2) by the position x_d at which the string is arrested by the finger. The point of greatest interest to us is that, in a string of length a which is arrested at the center, for example, all wave lengths a/m' are reduced in amplitude, where m' is an odd integer; or, in general, if the string is damped at some distance from the end expressed by a/g where g is any chosen integer, then the string cannot vibrate at any wave lengths except those which are expressed by $\lambda_p = 2a/gm''$, where m'' is any integer.

If it is assumed that the arresting of a vibrating string by placing a finger on that point is equivalent to loading the string at that point with a mass which is infinite in

comparison to the density of the string, then it becomes of interest to know what may happen if the added mass is finite. We shall not go into this matter, but fortunately this problem has been worked out,[17] and we shall merely state the results: (1) a mass of weight m attached to a string reduces the frequency of vibration of all standing waves except those for which the point of attachment is a node, and (2) the mass reduces the amplitude of vibration at the point of attachment. As the mass is increased, the amplitudes and frequencies decrease, until at infinite mass fully developed nodes are induced.

The two-dimensional standing waves, as observed on a rectangular drumhead,[18] are very convenient to study, because when such a drumhead is set into vibration by striking, sawdust scattered upon the membrane is observed to accumulate upon the nodal points or nodal lines. When such a drumhead is struck and then arrested at some point, other points are observed to be induced nodes, as evidenced by the accumulation of sawdust in small piles at these positions. Or, if the drumhead is arrested by placing a straight edge one-third of the way along the a dimension, then an induced nodal line is formed at $2a/3$.

Examples of Branching in Snow Crystals.—The theory which we wish to present is that snow crystals grow from the deposition of water molecules upon small nuclei of ice which are under thermal vibration corresponding to a temperature of 250°–273° K.; the molecules strike and bounce off the nuclei, with relatively few sticking; the number of those which strike and stay to contribute to the size of the growing crystal is greatly influenced by the frequency and amplitude of vibration at any point on the crystal; if a portion of a spine or branch of a dendrite becomes more heavily loaded than the surrounding portions, the vibrations are somewhat restricted, as is a string with a mass attached; these loaded portions become more loaded with further growth; therefore, the growth from the loaded point becomes a runaway process, and thus branching occurs; finally, the loaded portion simulates a node, and the induced nodes sponsor growth at other portions according to a defined pattern.

To clarify the process described in the above paragraph, let us discuss a number of the pictures from the book by Bentley and Humphreys.[10] First let us imagine that Figure 1 is an enlarged photograph of a very small nucleus which in the later stages of its growth has experienced "normal" growth for a sufficiently long time to have established a "normal" hexagonal shape. If conditions were to change so as to be conducive to dendritic growth,[20] then the six corners would find themselves receiving more molecules and losing more heat of crystallization than the flat portions. If this situation prevailed, the corners would "sprout" dendrites, as shown in Figure 2.

After a very much longer time the dendritic "sprouts" will have grown to quite extended lengths, as shown in Figure 3. Up to this point, the growth of the dendrites can be explained on the basis of the known theories.[20] To account for the branching, however, one must consider the standing-wave phenomena.

FIGS. 1–6.—Some typical snow crystals photographed by W. A. Bentley. (Reproduced by permission from *Snow Crystals*, by Bentley and Humphreys. Copyright 1931, McGraw-Hill Book Co.)

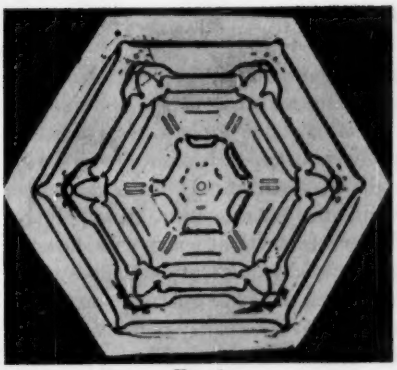


FIG. 1

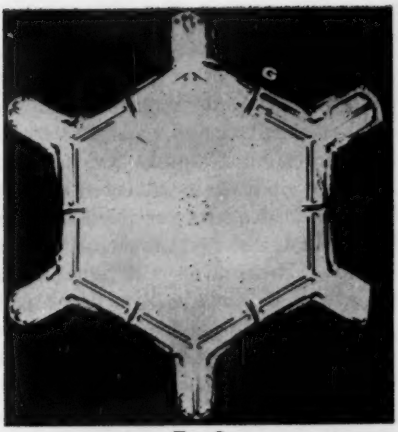


FIG. 2

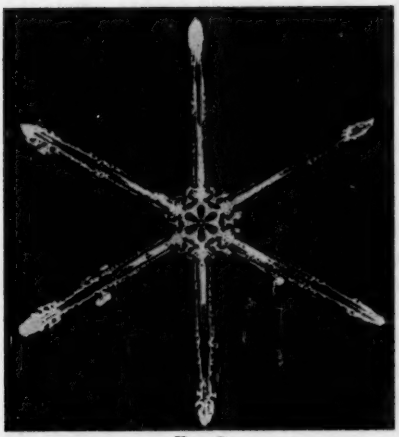


FIG. 3

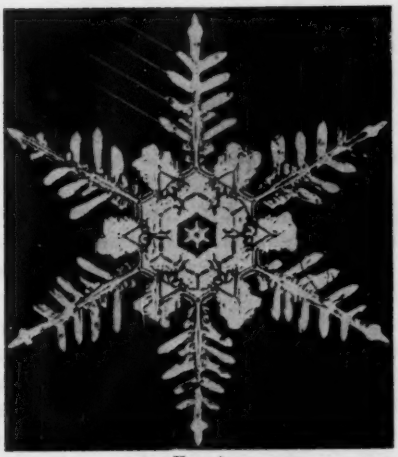


FIG. 4

FIG. 5

FIG. 6

Let us consider that each of the six spines in Figure 3 is an array of molecules[21] arranged in the order required by the structure of ice and that all the molecules are in thermal agitation in a manner demanded by the theories of black-body radiation and Planck's distribution law, but that the distribution of energies between the modes of vibration is influenced by the boundary conditions. If one of the spines becomes loaded at some point, say one-fourth the distance from their point of attachment, then nodes are induced at equal intervals along the spine. As shown in Figure 4, these nodes will "sprout" dendritic branches that are equally spaced. Figure 5 shows the equal spacing of the branches on a dendrite where the interval is one-eleventh of the spine length, and the nuclear size is extremely minute.

The question may arise as to how the waves which stand in one of the six branches are coupled with those in the other five branches. Experimentally one may build a model of six such spines from thin drill-rod material and weld them together, then place them on a thin dental rubber dam sheet. Sawdust may be used to study the wave forms produced by vibrating them. The torque through the point of intersection transmits the same frequencies and induces the same nodes in all branches in an identical manner.

Figure 6 is an interesting example of a crystal which changed its vibrational modes. This crystal grew dendritically from the center O to the point S according to some system of loading, then grew a slender spine from S to T, and then rebranched systematically again. Also, one may notice secondary branches such as the one marked B. Many of the secondary and tertiary branches have been inhibited by interference or shielding by their neighbors.

Returning to Figure 1, which was chosen for its simplicity, we see markings which indicate that it has formed complete hexagons at several stages in its development. Each time after (or during) the filling-in process, changing conditions produced further dendritic growth. Similar "filling-in" scars can be seen in Figure 2 at G'.

These markings on the interior of the crystals are almost as interesting as the branching characteristics of snow crystals. The complicated markings such as are shown in Figure 4 are a sort of a written record of the atmospheric conditions through which the crystal has gone in its relatively long experience of having been carried up and down repeatedly by upcurrents and gravity between elevations at which the temperature, humidity, and pressure differ extremely with altitude. A study of these markings should eventually become valuable to meteorologists. Nakaya[11] and his co-workers have established, in their experiments on the synthetic production of snow crystals, the conditions of ambient temperature, humidity, and supercooling under which each of the various types of snow crystals is formed. If it can be assumed that natural snow crystals of a given type can be used as evidence of the conditions in the air, then meteorology has a new tool.

In this discussion there is one point that has not been clarified. This point has to do with the question, "What is the mechanism by which some point on a dendritic spine becomes more heavily loaded with molecules than other points?" We shall consider this point statistically.

Some Statistics of Snow Formation.—Let us consider some facts which might enable us to imagine what goes on in the atmosphere during the growth of a snow crystal.

At 0° C., saturated air contains 4.835 gm. of water per cubic meter, according to the Smithsonian tables. At 20° C. the air is saturated when there is only 0.892 gm.

per cubic meter. This means that if saturated air at 0° C. is chilled to −20° C., about four-fifths of the suspended water molecules will precipitate as ice as fast as these molecules can establish (or find) nuclei upon which to deposit. Expressed in partial pressures, the 4.835 gm. of H_2O per cubic meter is about 0.006 atmosphere, or 4.56 mm. of mercury. This does not appear to be a very high pressure, but it represents 1.6×10^{17} molecules of H_2O per cubic centimeter mixed with about 200 times this many molecules of air. At this congestion, the water molecules are only 0.2×10^{-3} millimeter apart. They do not have far to go to associate with one another, and they should not require very long to accomplish much, since the H_2O molecules are traveling at a mean velocity of 62,700 cm/sec. For example, if an ice crystal is suspended in the saturated atmosphere at 0° C., it is bombarded with water molecules at the rate of 2.5×10^{21} molecules per square centimeter per second, that is, about $^1/_{250}$ mole/sec. This means that every 28 seconds there will be about 79 calories of heat deposited per square centimeter if every molecule sticks (this is assuming a heat of sublimation of 675 cal/mole). This accumulation of heat would raise the temperature and melt the crystal (79 cal/mole is the heat of fusion of ice).

To be more precise, let us consider an actual snow crystal of over-all dimensions of 100 μ or about $^1/_{10}$ mm. Assume that it is similar to the one pictured in Figure 3 having six spines, each 20 times as long as wide and about one-tenth as thick as wide. Such a crystal has an area of 10^{-6} square centimeter. Such an area is bombarded with 2.5×10^{15} water molecules per second. Using the figure 1.4×10^{15} and the volume occupied by water molecules in ice, one concludes that the crystal should grow at a rate of about one centimeter per second in all directions. This velocity of growth is much more rapid than observed rates,[6] since observed rates are approximately 100 μ in 30 seconds or 10^{-4} cm/sec. We are forced to conclude that even at fast rates of growth only about four out of every thousand molecules which strike the crystal surface stick; the other 996 bounce off.

As pointed out by private communication from Sterling B. Hendricks, the present argument applies not only to the sublimation processes but also to the surface migration of water molecules on ice. Murphy[22] has shown that in the bipedal random walk of water at 0° C. there are 10^5 skips along the surface for each jump off the surface, and at −100° C. the ratio is 10^8.

Now let us consider the vibrations within the ice crystal. The longest thermal vibration in the crystal has a wave-length twice the length of the spine or about $^1/_{10}$ mm, that is, in the far infrared, not far from the black-body maximum at 0° C.; and the shortest possible wave is twice the interatomic distance, or about 3 A, that is, in the X-ray region and negligibly weak in intensity. Assuming that the velocity of sound in ice is about 1,500 m/sec, we may calculate the lowest frequency of vibration of a dendritic spine to be 1.5×10^7 and for the highest frequency 6.6×10^{12} cycles/sec. This means that even for the lowest possible frequency of oscillation of the spine of an ice crystal only about 10^{-7} second is required to complete a wave cycle. Using the structure of ice, we may calculate the area of a unit cell of the crystal to be about 33×10^{-16} square centimeter. Following, in our imagination, the oscillation of a unit cell back and forth once every 10^{-7} second, and an exposed area of only 33×10^{-16} square centimeter, in an atmosphere where it is being bombarded at a rate of 2.5×10^{22} times per second per square centimeter with water molecules, we can calculate the number of bombardments per cycle per

unit cell to be approximately 8. Since only one of these out of each thousand sticks, we may conclude that, on the average, we must consider an expanse of about 300 unit cells to see one captured molecule per cycle at the slowest possible vibration rate. For higher frequencies the times are shorter and the expanse (over which one may expect an average of one capture per cycle) is greater.

This means that, while we may calculate the average number of molecules captured by a given area per second, we have no means of saying for certain that this really happens. Watching a single area in which we expect one molecule to be captured and confining our observations to the short time of one cycle, we have a good probability of seeing no molecules captured, or we have a fair probability of seeing two captured. According to our earlier study of vibrating strings, if two are captured, the system is loaded more heavily in this region than the average, and further captures are more likely on the next cycle. The runaway process has begun!

Fortunately, the probability of capturing any portion of the expected number of molecules is a problem in a field of statistics which has been worked out.[23] In calculating the deviations from expectation, Poisson developed an equation which has proved reliable in many problems such as production control, plant breeding, and other fields where data can be collected:

$$P(n) = \frac{\epsilon^n e^{-\epsilon}}{n!}, \tag{3}$$

where $P(n)$ represents the probability of n molecules falling on an area where ϵ molecules were expected to fall. This equation can be used to calculate the probability that any chosen area of a crystal may become abnormally overloaded with molecules during one vibration cycle. From a table of Poisson probability values[23] (p. 458), we see that, for $\epsilon = 1$, the probability of getting what we expect is about the same as the probability of getting none. The probability of getting 2 is about one-half that of getting 1, and the probability of getting 3 is about one-sixth. The maximum probability is always on the expected number; for example, when $\epsilon = 5$ the maximum probability is .175 for $n = 5$, and when $\epsilon = 10$ the maximum probability is .125 in the neighborhood of $n = 10$. The tables teach us that if we expect 1 molecule per unit area for a given cycle, we have a probability of about 1 out of 3 of getting it; if we expect 5 molecules, we have a probability of 1 out of 5 of getting it, and if we expected 10, our probability is 1 out of 8. The total probability is unity in any set, but the bulk of the probability is scattered over the many unexpected groupings of events.

This all means that, according to statistics, an uneven loading is very likely in a single vibration of a dendritic spine, and branching is easily initiated.

An interesting experiment would be to grow a piezoelectric crystal while vibrating, to see whether fins could be included at the nodes of the vibrations.

The author is indebted to the Stanford Research Institute and the Engineering Experiment Station of the University of Utah for facilities in carrying out this work. The author is also indebted to the McGraw-Hill Book Company, Inc., for permission to use selected figures from the book by Bentley and Humphreys.[10]

[1] I. Langmuir, *J. Meteorol.*, **5**, 175, 1948.

[2] I. Langmuir, *Proc. Am. Phil. Soc.*, **92**, No. 3, 167, 1948.

[3] V. J. Schaefer, *J. Meteorol.*, **6**, 283, 1949.

[4] V. J. Schaefer, *Chem. Revs.*, **44**, 291, 1949.

[5] V. J. Schaefer, General Electric Research Laboratory Occasional Report No. 11 (Schenectady, N.Y., 1948).

[6] V. J. Schaefer, *J. Appl. Math. and Phys.*, **1**, 153, 1950.

[7] William Scoresby, *An Account of the Arctic Regions, with a History and Description of the Northern Whale-Fishery* (Edinburgh, 1820).

[8] G. Hellmann, *Schneekrystalle* (Berlin, 1893).

[9] W. A. Bentley, *Monthly Weather Rev.*, **29**, 212, 1901; **30**, 607, 1902; **35**, 348, 397, 439, 512, 584, 1907; **46**, 359, 1918; **52**, 530, 1924; **55**, 358, 1927.

[10] W. A. Bentley and W. J. Humphreys, *Snow Crystals* (New York and London: McGraw-Hill Book Co., 1931).

[11] Ukichiro Nakaya, *Snow Crystals* (Cambridge, Mass.: Harvard University Press, 1954).

[12] J. D. Bernal and R. H. Fowler, *J. Chem. Phys.*, **1**, 515, 1933.

[13] H. D. Megaw, *Nature*, **134**, 900, 1934.

[14] E. F. Burton and W. F. Oliver, *Nature*, **135**, 505, 1935; *Proc. Roy. Soc. London, A*, **153**, 166, 1935.

[15] R. L. McFarlan, *J. Chem. Phys.*, **4**, 160, 253, 1936; *Phys. Rev.*, **49**, 199, 644, 1936.

[16] D'Arcy W. Thompson, *Growth and Form* (New York: Cambridge University Press, 1948).

[17] C. A. Coulson, *Waves* (London and New York: Interscience Publishers, 1947; 6th ed., 1952).

[18] William Elwood Byerly, *Fourier Series and Spherical Harmonics* (Boston: Ginn & Co., 1893), pp. 126–134.

[19] H. Morgenau and G. M. Murphy, *The Mathematics of Physics and Chemistry* (New York: D. Van Nostrand Co., Inc., 1943).

[20] The formation of spines reaching from a crystalline mass (or a single crystal) into the solvent has been explained (D. McLachlan, Allan Carlson, C. J. Christensen, and Alan King, University of Utah Engineering Experiment Station Bull. 57, 1952) on the assumption that under certain conditions the rate of deposition of molecules on the growing crystal is controlled by the rate at which any one of three things can occur: (1) the arrival of solute molecules to the various points on the surface (or the departure of solvent from the surface), (2) the escape of heat of crystallization from the surface, or (3) the escape of electric charges from the surface. The last of these processes is markedly exhibited by the needles formed in the fast electrodeposition of silver. It has been shown (*ibid.*) that all these processes can be expressed by a conformal mapping of the lines of flow and the equiconcentration contours about bodies of various shapes, and that protrusions on the crystal are favorably shaped to enhance any of the three processes and consequently develop spines.

[21] Leon Brillouin, *Wave Propagation in Periodic Structures* (New York: McGraw-Hill Book Co., 1946; 2d ed., New York: Dover Publications, 1953).

[22] E. J. Murphy, *J. Chem .Phys.*, **21**, 1831, 1953.

[23] Thornton C. Fry, *Probability and Its Engineering Uses* (New York: D. Van Nostrand Co., Inc., 1928), pp. 214–240, 458–462.

EXCITATIONS AND POLYMERIZATION

BY ALBERT SZENT-GYÖRGYI

INSTITUTE FOR MUSCLE RESEARCH, MARINE BIOLOGICAL LABORATORY, WOODS HOLE, MASSACHUSETTS

Communicated November 26, 1956

As described previously,[1] fluorescent substances, if frozen in water, show peculiar forms of long-lived electronic excitations. It has been suggested that the long lifetime of these excitations is due to their triplet state. It has also been shown

that various mono- or polyvalent alcohols interfere with these excitations, while other substances, like sugars, promote them.

It seems probable that dimerizations are involved in these phenomena, as described by Scheibe,[2] Rabinowitch and Epstein,[3] and others. If so, the excitations could be symbolized by the following equations:

$$\begin{aligned} &A + B \quad \rightarrow AB, \\ &h\nu + AB \rightarrow A^{**}B \rightarrow A^{*}B^{*}, \quad \text{or} \\ &h\nu + AB \rightarrow A^{*}B^{*}, \end{aligned}$$

A and B standing for the two dimerizing molecules and the asterisk for their excitation. Why dimerization facilities transition into the triple state has to be decided by further studies.

[1] A. Szent-Györgyi, *Science*, **124**, 873, 1956.
[2] G. Scheibe, *Z. angew. Chem.*, **50**, 217, 1937; *Z. Electrochem.*, **52**, 283, 1948.
[3] E. Rabinowitch and L. F. Epstein, *J. Am. Chem. Soc.*, **63**, 69, 1941.

---

*MALE SEGREGATION RATIO ADVANTAGE AS A FACTOR IN MAINTAINING LETHAL ALLELES IN WILD POPULATIONS OF HOUSE MICE**

BY DAVID BRUCK

COLUMBIA UNIVERSITY

Communicated by L. C. Dunn, November 23, 1956

It has been observed that certain populations of wild house mice (*Mus musculus*), comprising most of those sampled from different parts of the United States, contain heterozygotes for *t* alleles at locus *T*.[1] Since most of the alleles identified are recessive lethals,[2] the question arises as to what agencies are responsible for the high frequencies of lethals in wild populations.

One such agency has been identified by estimating the relative frequencies of gametes containing + and *t* alleles in the progenies from male heterozygotes (+/*t* and *T*/*t*). The ratio of *t* gametes to all gametes has been estimated from samples of several males each from twelve different populations. These ratios have been found to vary between 0.90 and 0.99; that is, *t* gametes are always transmitted by males to the great majority of offspring.[2] Transmission by female heterozygotes occurs in normal proportions (0.5).

This peculiarity (the cytogenetic mechanism for which is unknown) has been called "segregation ratio" abnormality.[3] Clearly this leads to a tendency for such alleles to establish themselves in populations by what has been termed "segregation ratio advantage." Prout[4] has examined, in a preliminary way, the relations between *m*, the male segregation ratio, lethality of *tt* homozygotes, and the gene frequency of *t* under random mating. He has shown that values of *m* exist sufficient to compensate for lethality and maintain the *t* allele at an equilibrium frequency; the conclusion was that segregation ratio advantage is an evolutionary force that must be evaluated.

Prout[4] obtained his formulas on the assumption that m was the same in both sexes. He then remarked that the same results would hold if the segregation ratio differed in the two sexes, provided that m was taken to be the arithmetic mean of the two values. In a personal communication Prout pointed out that this remark was incorrect and gave our formulas (3) and (5) for the case where only one sex has an anomalous segregation ratio. Since Prout is not publishing his results, these formulas are rederived below. Tables and graphs are presented, discussing this new derivation and its implications, including some properties not previously foreseen.

The Mathematical Model.—In seeking a simple model which relates male segregation ratio and gene frequency at equilibrium, the following assumptions are made:

1. Only two alleles + and t exist at this locus in the model population; + is completely dominant over t.
2. The adaptive values of the genotypes dealt with are $+/+ = 1$; $+/t = 1$; $t/t = 0$.
3. Mutation rates are negligible.
4. Mating is at random in an infinite reproductively isolated population.
5. The female segregation ratio is 0.5.

The terms used in deriving equilibrium values are as follows:

1. Parameters of the nth generation:
   q_n = frequency of t after selection (i.e., among adults),
   $1 - q_n$ = frequency of + after selection (i.e., among adults),
   D_n = frequency of $+/+$ before selection (i.e., before death of lethals),
   H_n = frequency of $+/t$ before selection (i.e., before death of lethals),
   R_n = frequency of t/t before selection (i.e., before death of lethals).
2. Male segregation ratio:
$$m = \frac{\text{Number of gametes transmitting } t}{\text{Total number of functioning gametes from each male heterozygote.}}$$
3. Frequency of $+/t$ heterozygotes after selection, $a = 2q_n$.
   Frequency of $+/+$ homozygotes after selection, $b = 1 - 2q_n$.

It follows that

$$a = 2q_n = \frac{H_n}{D_n + H_n}. \tag{1}$$

It should be noted that in the present paper q_n denotes the gene frequency after selection, whereas in Prout [4] q_n denotes the frequency before selection.

Since the gametic frequencies for a male heterozygote are m (t) to $(1 - m)$ (+), while those for females are one-half (t) to one-half (+), we see by Table 1 that the progeny of a mating of two heterozygotes are in the ratio of $(1 - m)/2$ $(+/+)$ to $(1 - m)/2 + m/2$ $(+/t)$ to $m/2$ (t/t). Outcomes of other matings can be obtained similarly.

TABLE 1

Gametes of Male	Gametes of Female: +	Gametes of Female: t	Total
+	$(1 - m)/2$	$(1 - m)/2$	$1 - m$
t	$m/2$	$m/2$	m
Total	$1/2$	$1/2$	1

With random mating the frequency of the three genotypes in the $(n + 1)$th generation before selection will be as given in Table 2.

TABLE 2

Type of Mating (♀ × ♂)	Frequency of Type of Mating	Frequency of Genotypes in $(n + 1)$th Generation before Selection		
		+/+	+/t	t/t
+/t × +/t	a^2	$a^2(1 - m)/2$	$a^2/2$	$a^2m/2$
+/t × +/+	ab	$ab/2$	$ab/2$	...
+/+ × +/t	ab	$ab(1 - m)$	abm	...
+/+ × +/+	b^2	b^2	...	...

Adding the results from Table 2, we have

$$D_{n+1} = b^2 + \frac{a^2(1 - m)}{2} + \frac{ab(3 - 2m)}{2},$$

$$H_{n+1} = \frac{a^2}{2} + \frac{ab(2m + 1)}{2}, \tag{2}$$

and

$$R_{n+1} = \frac{a^2m}{2}.$$

But from equation (1) we have

$$q_{n+1} = \frac{H_{n+1}}{2(D_{n+1} + H_{n+1})},$$

and since equations (2) give H_{n+1} and D_{n+1} in terms of a and b, which in turn are equal to $2q_n$ and $1 - 2q_n$, respectively, we can express q_{n+1} in terms of q_n and m as

$$q_{n+1} = \frac{q_n(2m - 4mq_n + 1)}{2(1 - 2mq_n^2)}. \tag{3}$$

If we now set $q_{n+1} = q_n$, we arrive at the equilibrium formula:

$$\hat{q} = {}^1/_2 \pm \frac{\sqrt{m(1 - m)}}{2m}. \tag{4}$$

Considering the solution $\hat{q} = {}^1/_2 + \sqrt{m(1 - m)}/2m$, it is clear that such a solution is not consistent with reality, since, no matter what value m takes, $\hat{q}$ is greater than 0.5, and the frequency of $+/t$ heterozygotes after selection, or $2\hat{q}$, is greater than 1.0, which is impossible.

It can be easily demonstrated that the other solution,

$$\hat{q} = {}^1/_2 - \frac{\sqrt{m(1 - m)}}{2m}, \tag{5}$$

holds if and only if m is greater than 0.5, i.e., in order for $\hat{q}$ to be positive, $\sqrt{m(1 - m)}/m$ must be less than 1, and this is true if and only if m is greater than 0.5, since, if

$$\frac{\sqrt{m(1 - m)}}{m} < 1,$$

then

$$\frac{1-m}{m} < 1,$$

and

$$\frac{1}{m} - 1 < 1,$$

or

$$\frac{1}{m} < 2, \qquad m > 0.5.$$

Furthermore, it is clear that any consecutive pair of steps imply each other, so that, if $m > 0.5$, then $\sqrt{m(1-m)}/m < 1$. It is clear that if m is less than 0.5, $\hat{q}$ is 0. It follows from equations (5) and (2) that the genotypic frequencies before selection at equilibrium are

$$\begin{aligned} \hat{D} &= 1 - (1+2m)\hat{q} - 2m\hat{q}^2, \\ \hat{H} &= \hat{q}(1+2m) - 4m\hat{q}^2, \\ \hat{R} &= 2m\hat{q}^2. \end{aligned} \tag{6}$$

If we consider $\hat{q}$, $2\hat{q}$, $1 - 2\hat{q}$ $\hat{D}$, $\hat{H}$, $\hat{R}$, and as functions of m (Fig. 1 and Tables 3 and 4), we note an expected rise in $\hat{q}$ and $2\hat{q}$ from a minimum of 0 to a maximum of

TABLE 3

m	$\hat{q}$	$2\hat{q}$	$1-2\hat{q}$	$\hat{D}$	$\hat{R}$
0.00	0	0	1	1	0
0.10	0	0	1	1	0
0.20	0	0	1	1	0
0.30	0	0	1	1	0
0.40	0	0	1	1	0
0.50	0	0	1	1	0
0.55	0.048	0.095	0.905	0.903	0.003
0.60	.092	0.184	.816	.808	.010
0.65	.133	0.266	.734	.717	.023
0.70	.173	0.345	.655	.637	.042
0.75	.211	0.423	.577	.539	.067
0.80	.250	0.500	.500	.450	.100
0.85	.290	0.580	.420	.360	.143
0.90	.33	0.667	.333	.267	.200
0.95	.385	0.771	.229	.165	.282
1.00	0.500	1.000	0.000	0.000	0.500

TABLE 4

m	$\hat{H}$	m	$\hat{H}$	m	$\hat{H}$	m	$\hat{H}$
0.00	0.000	0.50	0.000	0.82	0.470	0.92	0.544
.10	.000	.55	.095	.84	.488	0.94	.551
.20	.000	.60	.182	.86	.505	0.96	.554
.30	.000	.65	.260	.88	.520	0.98	.547
0.40	0.000	0.70	0.331	0.90	0.533	1.00	0.500
		.75	.394				
		.80	.450				

0.5 and 1.0, respectively, and the corresponding expected decrease of $1 - 2\hat{q}$. We would intuitively expect the corresponding rise in $\hat{R}$, and decrease in $\hat{D}$, but the rise

of $\hat{H}$ to a maximum and its subsequent decline to 0.5 at $m = 1$ are somewhat unexpected.

A vague but intuitive way of viewing this behavior of $\hat{H}$ is to note that $\hat{H}$ depends on the frequency of both t and $+$ and that ratios lower than this would produce too few t alleles and higher ratios would produce too few $+$ alleles for a maximum $\hat{H}$.

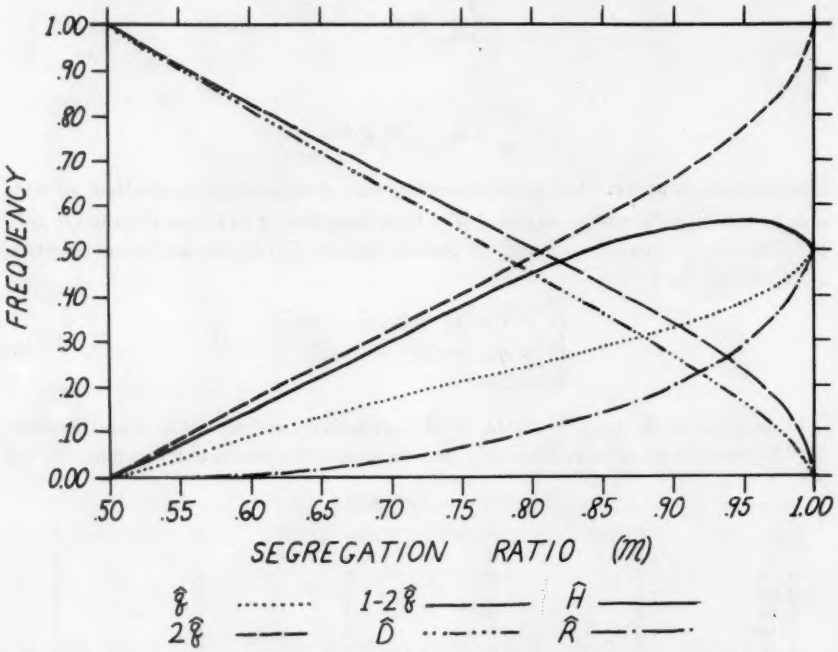


FIG. 1.—The equilibrium values of $\hat{q}$, $2\hat{q}$, $1 - 2\hat{q}$, $\hat{D}$, $\hat{H}$, and $\hat{R}$, are plotted against m, the segregation ratio. Numerical values are given in Tables 3 and 4.

The value of m for which $\hat{H}$ is maximum was obtained by setting

$$\frac{d\hat{H}}{dm} = 1 + \frac{2 - 2m - {}^1/_2 m}{2\sqrt{m(1-m)}} + \frac{\sqrt{m(1-m)}}{2m^2} = 0$$

and solving for m. $\hat{H}$ is found to attain its maximum value, viz., 0.554, when m is 0.959, as shown in Table 4 and Figure 1. It is a curious fact that this value of m is close to the average value of m found experimentally by testing males heterozygous for each of thirteen t-alleles each derived from a different wild ancestor. The average value of the segregation ratios was about 0.96.[5]

It was seen in plotting the frequency of t against generations for $m = 0.5$, 0.6, 0.7, 0.8, 0.9, and 1.0, with an arbitrary set of starting frequencies of $+/+$ and $+/t$ before selection, that the equilibrium value of t was approached to within 2 per cent by the sixtieth generation. Figure 2 illustrates this, employing $m = 0.9$.

Summary.—Genes such as the t alleles in house mice, which are lethal when homozygous and produce abnormal segregation ratios in male heterozygotes, are

considered from the point of view of their population dynamics. Prout has given formulas which he erroneously stated apply to this situation. In correspondence he has pointed out the error and given the correct formulas, which is rederived in the present paper. These formulas hold for all values of m, the proportion of functional male gametes from $+/t$ males. For $m > 1/2$, t reached an equilibrium frequency at $q = 1/2 - \sqrt{m(1-m)}/2m$ in adults.

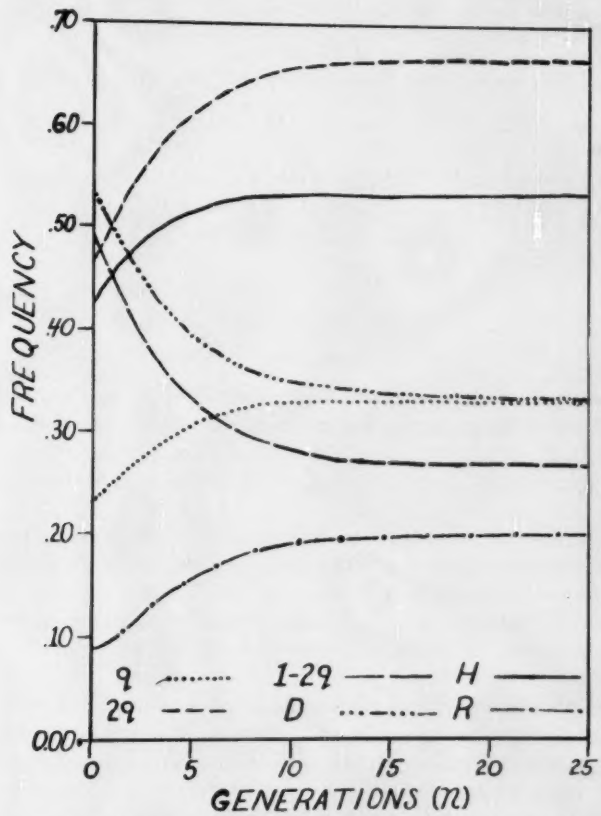


FIG. 2.—Values of q, $2q$, $1 - 2q$, D, H, and R are plotted as they change for each successive generation, starting with the arbitrary set of initial values $D = 0.49$, $H = 0.42$, $R = 0.09$, for $m = 0.9$. The graph illustrates the rapid approach of all variables toward their equilibrium values, even if the starting values differ markedly from the latter.

Tables and graphs of q as a function of time and of the equilibrium values of various gene and genotypic frequencies both before and after selection are given as functions of m, and agreement of these with what one might expect is apparent except in the case of the frequency of heterozygotes $+/t$ before the death of lethal zygotes, which reaches a maximum at $m = 0.959$.

Grateful thanks are due to Dr. Timothy Prout, of the University of California, who introduced me to this problem, to Professor L. C. Dunn, under whose guidance this paper was written, and to Professor Howard Levene, who checked the consistency of the model and made many helpful suggestions.

* Work performed in the Department of Zoölogy and Nevis Biological Station under Contract AT(30-1)-1804 with the United States Atomic Energy Commission.

[1] L. C. Dunn, *Nature*, **176**, 1275–1276, 1955.
[2] L. C. Dunn, these PROCEEDINGS, **43**, 158–163, 1957.
[3] L. C. Dunn, *Acta genet. et statist. med.*, **4**, 139–147, 1953.
[4] T. Prout, *Acta genet. et statist. med.*, **4**, 148–151, 1953.
[5] L. C. Dunn, *Cold Spring Harbor Symposia Quant. Biol.*, Vol. **21**, 1956 (in press).

---

*EVIDENCE OF EVOLUTIONARY FORCES LEADING TO THE SPREAD OF LETHAL GENES IN WILD POPULATIONS OF HOUSE MICE**

BY L. C. DUNN†

COLUMBIA UNIVERSITY

Communicated November 23, 1956

A population changes in an evolutionary sense when there is a shift prevailingly in one direction of the proportions of genotypes of which it is composed. It is for this reason that agencies tending to change these proportions, derived primarily from the relative frequencies of gene forms or alleles, are referred to as "evolutionary forces."

Genotypic diversity, the usual condition of all cross-fertilizing populations in nature, tends to remain constant unless the gene frequencies are altered by one or more of these evolutionary forces. For any one pair of alleles, such as A and a, with frequencies q and $1 - q$, the constant or equilibrium frequencies of genotypes under random mating have long been known to be of the form $q^2AA + 2q(1 - q)Aa + (1 - q)^2aa = 1$.

Agencies hitherto recognized as leading to alteration of the equilibrium frequencies are (1) mutation pressure, changes from A to a and a to A at unequal rates; (2) natural selection, that is, unequal fitness or adaptive values of the several genotypes, such as lethality or lesser fertility of AA, Aa, or aa, relative to the others; (3) random fluctuation of gene frequencies in small sections of the population (random genetic drift); (4) differential migration of genotypes into or out of the population.

Equality of the genotypes in all these respects is of course the condition for maintenance of constant proportions.

The main purpose of this report is to present some new facts indicating that a fifth evolutionary force is primarily responsible for the maintenance of high frequencies of certain lethal genes in populations of wild house mice and to examine the effect of this force on equilibrium frequencies of such lethals. This fifth force I refer to as segregation ratio, since it alters the basic Mendelian proportion of functional gametes produced by heterozygotes Aa. In all the above equilibrium com-

putations this ratio is assumed to be 0.5; actually, in the cases I shall present, it averages in one sex (males) 0.95 to 0.05, in favor of the gametes transmitting the lethal, leading to an enormous advantage of a gene against which complete selection in homozygous condition is exercised, since all such homozygotes are eliminated before birth. The gene is thus enabled to spread against what might be called "evolutionary gravity."

The facts are as follows. Many, probably most, wild populations of the house mouse—*Mus musculus*—in the United States, living either near dwellings or in nature, contain animals heterozygous for variant alleles at one locus *T*. Such heterozygotes are detected by bringing wild mice into the laboratory and mating them with animals containing a dominant marker allele *T* which by itself produces a short tail. Wild mice containing a recessive at this locus produce in such crosses a unique class of offspring—tailless—which, when bred together, give rise to a true-breeding tailless line of genotype T/t^w, t^w being a recessive lethal allele of *T* (from a wild heterozygote) which interacts with *T* to produce taillessness. In a few cases t^w alleles have proved to form viable homozygotes t^wt^w, which, however, are male-sterile, so these, too, have an adaptive value of zero in an evolutionary sense.

All tailless lines with a lethal *t*-allele maintain themselves as balanced lethal systems with litter sizes reduced by death of two classes of lethal homozygotes to half the normal size.[1] Seventeen such balanced lethal lines have been studied, each originating from a single heterozygote from a wild population. Data from fourteen such lines are shown in Table 1. The second and seventh lines represent the be-

TABLE 1

BALANCED LINES MAINTAINED BY MATINGS OF TAILLESS BY TAILLESS ($T/t \times T/t$)

(Each Allele Derived from a Single Wild Heterozygote $+/t$)

ALLELE	POPULATION	OFFSPRING		LITTER SIZE
		Tailless	Normal	
t^{w1}	N.Y. 1	399	13*	3.3
t^{w2}	N.Y. 1	355	235	5.7
t^{w3}	Conn. 1	438	5*	3.5
t^{w5}	N.Y. 2	605	2	3.6
t^{w6}	Fla. 2	345	..	3.2
t^{w7}	Tex. 1	75	71	4.9
t^{w11}	Conn. 2	521	..	3.6
t^{w12}	Calif. 1	418	1*	3.9
t^{w13}	Tex. 2	161	..	3.6
t^{w14}	Vt. 1	183	1	3.7
t^{w15}	Kan. 2	92	1	3.7
t^{w16}	Vt. 2	71	1	3.9
t^{w17}	Ariz. 1	59	..	3.0
t^{w18}	From t^{w11}	92	..	4.0

* Dead at birth; microcephalic.

havior of viable alleles; the other twelve lines, lethal alleles. The exceptional normal-tailed offspring in the first, third, and twelfth lines, which were dead at birth, represent survivors of the lethal homozygotes which die between the eleventh day and birth. The other exceptions are probably mutants in each of which a lethal t^w allele—early lethals in these lines—has mutated to a different t^w allele. This has proved to be the explanation in all cases analyzed. This, then, is the usual behavior of balanced lethal lines containing variant alleles at locus *T*, occurring in different wild populations.

We have now to ask why such alleles, most of them lethal, are maintained in wild populations. Lethals should be eliminated by natural selection. What preserves these lethals in high frequency?

Our first answer came from testing the ratios in which the two alleles + and t^w or T and t^w are transmitted to the next generation. The transmission ratio in females proved to be normal, i.e., 0.5 of each allele. Male heterozygotes, however, transmit the variant t allele to a great majority of offspring. The clearest proof of this is given in Table 2. Individual males T/t^w were tested by normal females known to be of genotype $+/+$. From 90 to 99 per cent of the offspring were of genotype $+/t^w$ (phenotypically normal); from 10 to 1 per cent were of genotype $+/T$ (phenotypically short-tailed). Similar ratios were obtained from $+/t^w$ males.

TABLE 2

RESULTS OF TESTING SEGREGATION RATIOS OF TAILLESS MALES WITH ALLELES FROM DIFFERENT WILD POPULATIONS

(Progenies from T/t^w♂ × $+/+$ ♀♀)

		Offspring			Ratio
Allele	From Population	Normal $+/t^w$	Brachy $+/T$	Total	$\frac{tw}{\text{Total}}$
t^{w1}	N.Y. 1	325	38	363	0.89
t^{w2}	N.Y. 1	319	15	334	.95
t^{w3}	Conn. 1	383	3	386	.99
t^{w4}	Wis. 3	187	5	192	.97
t^{w5}	N.Y. 2	479	30	509	.94
t^{w6}	Fla. 2	230	4	234	.98
t^{w7}	Tex. 1	214	16	230	.93
t^{w11}	Conn. 2	518	19	537	.96
t^{w12}	Calif. 1	393	18	411	.96
t^{w13}	Tex. 2	502	11	513	.98
t^{w14}	Vt. 1	392	20	412	.95
t^{w15}	Kan. 2	391	40	431	.91
t^{w16}	Vt. 2	107	6	113	0.95

One of the striking features of these tests is the similarity of the transmission ratio in all alleles derived from the wild, which, from evidence I cannot cite here in detail, we know to contain a variety of different, nonidentical alleles. From previously published work[2] we know that different alleles, arising in laboratory stocks and not exposed to selection in nature, show in males a variety of departures from the usual segregation ratio in both directions. Two salient facts in which wild populations agree are therefore (1) presence of lethals at this locus and (2) asymmetrical transmission ratios always favoring the lethal alleles.

We turn now to examine the theoretical consequences of this latter peculiarity. A preliminary account was given in 1953.[3] By the construction of mathematical models (due in the first instance to Prout[3] and corrected and extended by David Bruck[4] in our laboratory) it is found that values of the segregation ratio exist sufficient not only to compensate for the decrement in gene frequency of t due to death of lethal homozygotes but also to bring the frequency of t to its theoretical limit of 50 per cent, that is, to the condition in which the whole population, at equilibrium, is heterozygous for a lethal. Bruck's[4] Figure 1 (p. 156) shows the relation at equilibrium between m, the segregation ratio in males, and gene frequency, t, under the conditions of (1) absence of mutation pressure, (2) random mating in an infinite closed population, (3) complete selection against lethal homo-

zygotes and no selection for or against heterozygotes. It is obvious from these relations, as it was a priori, that the frequency of t is a function of m. The frequency of $+/t$, before selection against tt (death of lethal homozygotes), reaches a curious maximum above 50 per cent, at about 56 per cent, at a value of $m = 0.96$, which is close to the observed average value of m for the thirteen wild alleles tested. The biological meaning of this is not evident, and it may be coincidence. The frequencies of $+/t$ computed *after* selection reach a maximum at 100 per cent, the highest incidence attainable by a heterozygote for a lethal. The effect of the segregation ratio advantage by itself is sufficient to effect fixation, in the above sense, of a lethal in a wild population in which such an allele has occurred by mutation or has been introduced by migration.

We have begun experimental tests of the degree to which one of the assumptions on which the above equilibrium computations rest corresponds to conditions found in nature. Is it true that homozygous normals ($+/+$) and heterozygotes ($+/t$) are equivalent in fitness, i.e., have equal survival value? Starting with a population of wild mice in which the proportions of $+/+$ and $+/t$ are equal at birth, we find that at the time of major reproductive activity (from the end of the second to the end of the sixth month of life) such a population actually consists of about two-thirds $+/t$ to one-third $+/+$. The experimental figures are at present 42 $+/t$, 20 $+/+$, the deviation from equality being significant at the 0.05 level. Most of the deficiency of $+/+$ is probably due to early differential mortality. In a previous test[5] we found more sterile males among the $+/+$ class. The indications are that $+/t$ is superior in early life in viability and fertility. We have still to determine the total net contributions of the two genotypes to the next generation. If in this also the heterozygotes are superior, we shall conclude that the assumption of adaptive equality of the two surviving genotypes is invalid and shall have to correct for this in computing gene frequencies at equilibrium. Since it is probable that t alleles are favored both by transmission ratio and by natural selection, both tending to increase the frequency and spread of lethal alleles, we shall have to seek for other agencies tending to restrain them and keep the populations in balance. Our attention is now turning to one of the other assumptions on which our present equilibrium computations are based, namely, unlimited population size. If the effective breeding sizes of populations in nature are very small, then we shall have inbreeding, with consequent increase in relative frequency of homozygotes. Since one homozygous class is lethal, this would tend to decrease the frequency of such alleles. The theoretical model for this has not yet been worked out, but a priori one would not expect the intensity of continuous inbreeding which would be required to produce effects comparable in magnitude to the positive effects of transmission ratio or of selection favoring heterozygotes.

We have also some preliminary data[6] indicating that mutation from $+$ to t is not sufficiently frequent to generate a significant pressure toward increase of frequency of such alleles. The fourth parameter, differential migration, has not been measured at all.

Discussion.—In summary then, we may say that, of the evolutionary forces identified as acting to change the frequency of alleles at this one locus in nature, the newly recognized force—segregation ratio—is quantitatively the most important, with heterozygote advantage probably playing a secondary role but not

yet quantitatively evaluated. This leads to the paradoxical situation that, in this species, genes leading to death and sterility are not only maintained but encouraged. This looks so much like race suicide that it is difficult for me to think that it will be found frequently.

In one other case, in *Drosophila*, studied by Sturtevant and Dobzhansky[7] in 1936, a meiotic irregularity in spermatogenesis was found to be responsible for the transmission of a sex-linked mutation "sex-ratio" to the great majority of offspring. In males carrying this mutation the X-chromosome divides twice at meiosis, while the Y-chromosome is usually eliminated, so that nearly all sperm carry an X-chromosome with the mutation and thus produce only XX daughters, with occasionally an XO son. The great excess of mutant sperm should lead to a rapid spread toward fixation of the mutant in the population, which would then consist almost entirely of females and thus bring itself to an end. However, in this case, as Wallace[8] has shown, females homozygous for the mutant and males carrying it are less fit than animals lacking the mutant, so that its frequency in a confined population declines when in competition with the normal allele. Race suicide is thus prevented by selection against the mutant which prevents its spread. In the case I have described, however, the spread of the lethal is facilitated both by segregation ratio and by heterozygote advantage. I can only conclude that populations in which such mutants are present—and they seem to occur in most—must adjust themselves to the lethals by means at present unknown. Search for these means may reveal new facts about the structure of wild mouse populations.

The present case was discovered because of the extreme inequality in the transmission ratio in one sex. Meiotic irregularities connected with other complex loci in man and other forms might well result in changes of lesser degree in transmission ratio and would escape notice unless deliberately sought for. Only after this has been done can the importance of this new evolutionary force be evaluated.

Summary.—An evolutionary force, additional to those usually considered effective in altering gene frequencies (mutation, selection, random genetic drift and migration), is found to play the dominant role in determining the frequency of lethal alleles at locus *T* in populations of wild house mice. The new force, *male segregation ratio*, changes the normal ratio from 0.5 to an average of 0.96 in favor of the mutant allele. Thirteen alleles from twelve different wild populations from different parts of the United States all show similar segregation ratios. This advantage favors the maintenance and spread of lethal alleles. In addition, preliminary evidence indicates that animals heterozygous for such alleles have a selective advantage over those of normal genotype during early reproductive life. The effects of these facts on equilibrium values of such alleles in wild populations are discussed.

* For presentation at the November, 1956, meeting of the Academy.

† Studies of wild populations have been carried out at Nevis Biological Station, Irvington, New York, since June 1, 1955, under Contract AT(30-1) 1804 with the United States Atomic Energy Commission.

[1] L. C. Dunn and J. Suckling, *Genetics*, **41**, 344–352, 1956.

[2] L. C. Dunn and S. Gluecksohn-Waelsch, *Genetics*, **38**, 261–271, 1953.

[3] L. C. Dunn, with an appendix by T. Prout, *Acta genet. et statist. med.*, **4**, 139–151, 1953.

[4] D. Bruck, these PROCEEDINGS, **43**, 152–158, 1957.

[5] L. C. Dunn and J. Suckling, *Am. Naturalist*, **89**, 231–233, 1955.

[6] L. C. Dunn, *Cold Spring Harbor Symposia Quant. Biol.*, Vol. 21, 1956 (in press).
[7] A. H. Sturtevant and T. Dobzhansky, *Genetics*, 21, 473–490, 1936.
[8] B. Wallace, *Evolution*, 2, 189–217, 1948.

COMBINATORIAL PROBLEMS IN THE THEORY OF GRAPHS. IV

BY G. W. FORD AND G. E. UHLENBECK

INSTITUTE FOR ADVANCED STUDY, PRINCETON, NEW JERSEY,
AND DEPARTMENT OF PHYSICS, UNIVERSITY OF MICHIGAN

Communicated August 20, 1956

1. *Introduction.*—In the previous paper of this series[1] we derived asymptotic expressions for the numbers of star trees of a large number of points which are built out of stars from a given collection of types of star. These expressions were obtained from the functional equations satisfied by the counting series, as derived in Papers I and II. The only essential restriction in the derivation was the restriction that the collection of types of star be *finite*. With this restriction we could show that the counting series was convergent and that the singularities on the circle of convergence were all branch points of order 2. From this knowledge of the analytic behavior of the counting series, the asymptotic expressions for the coefficients could then be obtained, using standard methods.

The question arises whether the restriction to a finite collection of types of star is necessary. If *no* restriction is required, then our results would also be valid for the general connected graph of p points. In the next section we will review some results of Riddell and Polya[2] on the asymptotic numbers of graphs, connected graphs, and stars of p points and k lines, from which it is apparent that the counting series for the general connected graph of p points is always divergent. Clearly, then, *some* restriction on the collection of types of star is necessary for the convergence of the counting series. In Section 3 we will show that the results of Paper III are still valid for an infinite collection of types of star, if only each star in the collection has no more than a fixed number of independent cycles.[3] This restriction on the collection of types of star is surely not necessary for the validity of the results of Paper III. Whether necessary conditions can be formulated is an open question.

2. *The Asymptotic Results of Riddell and Polya.*—Riddell has investigated the asymptotic behavior of the numbers $C(p, k)$ and $S(p, k)$ of connected graphs and of stars with p labeled points and k lines. His result may be summarized by the following statement: For large p and k the overwhelming majority of all graphs are stars. More precisely, from equation (I, 18) and from equations (I, 20) and (I, 23) Riddell obtains the expansions

$$C(p, k) \sim \binom{{}^1\!/_2 p(p-1)}{k} - p\binom{{}^1\!/_2 (p-1)(p-2)}{k} + \ldots, \tag{1}$$

$$S(p, k) \sim \binom{{}^1\!/_2 p(p-1)}{k} - p\binom{{}^1\!/_2 (p-1)(p-2)}{k} - p(p-1)\binom{{}^1\!/_2 (p-1)(p-2)}{k-1} + \ldots. \tag{2}$$

For $k > p \log p$ one can readily show that the correction terms are dominated by the leading term in both these expressions.[4] This term is just the total number of graphs with p labeled points and k lines.

The corresponding result of Polya refers to the number $\pi(p, k)$ of graphs of p unlabeled points and k lines and may be summarized by the following statement: For large p and for $0 \ll k \ll {}^1/_2p(p-1)$ the overwhelming majority of graphs have symmetry number equal to unity. More precisely, from expression (II, 21), Polya obtains

$$\pi(p, k) \sim \frac{1}{p!} \binom{{}^1/_2p(p-1)}{k}, \tag{3}$$

if p is large and if k is not near the two ends of its range.[5] The right-hand side is seen to be the total number of graphs with p *labeled* points and k lines divided by $p!$, and since the number of graphs with p labeled points corresponding to a given graph of p unlabeled points is $p!$ divided by the symmetry number of the graph, it follows that the symmetry number of the overwhelming majority of graphs is unity.

From these results it follows that for large p the total number of connected graphs with p labeled points is asymptotically

$$\sum_k \binom{{}^1/_2p(p-1)}{k} = 2^{p(p-1)/2},$$

while in the case of unlabeled points one must divide this number by $p!$. In *both* cases the counting series according to the number of points is clearly always divergent.

3. *Conditions for the Convergence of the Star-Tree Counting Series.*—We will restrict ourselves in this section to the case of star trees with labeled points, since the discussion for the case of unlabeled points, although considerably more complicated, would be entirely similar. In Paper I we showed that if

$$t(x) = \sum_{p=1}^{\infty} t(p) \frac{x^p}{p!}, \tag{4}$$

where $t(p)$ is the number of star trees with p labeled points which are built out of stars from a given collection of types of star, then $T(x) = x\, dt/dx$ satisfies the implicit functional equation

$$F(T, x) \equiv T - x \exp\{S'(T)\} = 0 \tag{5}$$

in which $S'(T) = dS/dT$, and

$$S(z) = \sum_\sigma \frac{1}{s_\sigma} z_\sigma. \tag{6}$$

In equation (6) the sum is over all types of star in the collection, and s_σ and q_σ are, respectively, the symmetry number and the number of points in a star of type σ.

The method used in Paper III depended on the assumption that the function $S(z)$ was regular in a sufficiently large region around $z = 0$. We insured this by making the collection of types of star finite, so that $S(z)$ is a polynomial. From

the regularity of $S(z)$ around $z = 0$, it follows that $F(T, x)$ is regular in x and T in a certain region around $x = 0$ and $T = 0$. Since $\partial F/\partial T \neq 0$ for $x = 0$ and $T = 0$, it then follows from the implicit function theorem that $F(T, x) = 0$ may be solved for T as an analytic function of x in some region around $x = 0$. The singularities of $T(x)$ occur at those values of x and T for which $F = 0$ and

$$\frac{\partial F}{\partial T} = 1 - TS'(T) = 0. \tag{7}$$

Since $S(z)$ is a power series with positive coefficients, the nearest singularity z_0 of $S(z)$ must lie on the positive real axis. If z_0 is larger than T_0, which is the smallest root of equation (7), then, since T_0 is clearly positive, the value of x corresponding to T_0, denoted by x_0, will be the radius of convergence of $T(x)$.[6] In Paper III we showed that, after determining x_0 and $T(x_0) = T_0$ from equations (7) and (5), one could obtain the expansion of $t(x)$ around $x = x_0$, from which asymptotic expressions for $t(p)$ follow.

To construct an infinite star collection for which $z_0 > T_0$, so that the results of Paper III would be applicable, consider all stars with no more than c independent cycles. Any star with a given number of cycles will be one of a *finite* number of homeomorphic types.[7] For instance, all stars with two cycles consist of three chains[8] connected at their ends; all stars with three cycles belong to one of the four homeomorphic types shown in Figure 1, in which the curved lines represent chains and the points indicated are the *principal points*[8] of the homeomorphic type.

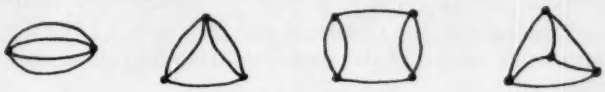


Fig. 1

In the next section we will show that the $S(z)$ corresponding to the infinite star collection of all stars of a given homeomorphic type has radius of convergence 1. Since there are only a finite number of homeomorphic types with no more than c independent cycles, the corresponding $S(z)$ will also have the radius of convergence 1. In Paper III we saw that for the case $c = 1$ (mixed Husimi trees) $T_0 = 0.456$. . . . For the larger collection of stars T_0 is therefore surely less than 1, and hence all the conclusions of Paper III remain valid.

4. *The Number of Stars of a Given Homeomorphic Type.*—Consider a given homeomorphic type which has K chains and P principal points. Let S be the symmetry number of the homeomorphic type,[9] and let there be L_1 single chains joining pairs of principal points, L_2 double chains, etc. The number of stars with p labeled points which are of this homeomorphic type we denote by $S(p)$.

The determination of $S(p)$ is a straightforward combinatorial problem. One has to multiply the number of ways of selecting P points with the number of ways these P points can be assigned to the principal points; the resulting number must then be multiplied by the number of ways the remaining p-P points can be distributed over the K chains to form a star, taking into account the symmetries produced by multiple chains. The result is

$$S(p) = \binom{p}{P} \cdot \frac{P!}{S} \cdot \frac{(p-P)!}{\prod_{m=1}^{K} (m!)^{L_m}} \sum_{[n_i]}' 1, \tag{8}$$

where the round sum goes over all sets of numbers $n_1, n_2, \ldots, n_K$, with the two conditions (indicated by the primes) that $\sum n_i = p - P$ and that in the case of multiple chains no more than one of the corresponding n_i can be zero.

If we form the counting series

$$S(z) = \sum_{p=P}^{\infty} \frac{1}{p!} S(p) z^p, \tag{9}$$

then, introducing equation (8), one easily obtains

$$S(z) = \frac{z^P}{S} \prod_{m=1}^{K} \left[\frac{z^m + m z^{m-1}(1-z)}{m!(1-z^m)} \right]^{L_m}. \tag{10}$$

Since for any star collection the $S(z)$ as defined by equation (6) is identical with the $S(z)$ as defined by equation (9),[10] one sees from the form (10) that for the collection of all stars of a given homeomorphic type the $S(z)$ has the convergence radius 1.

5. *The Average Number of Symmetries in Star Trees.*—We noted in Section 2 the Polya result that for large p and k the overwhelming majority of all connected graphs have symmetry number 1. The question naturally arises as to whether this is also true for those connected graphs which are star trees built out of stars with no more than c cycles, as considered in Section 3. Clearly this question can be answered by comparing $t(p)$, the number of star trees with p labeled points, with $\theta(p)$, the corresponding number of star trees with unlabeled points, since

$$\frac{t(p)}{p!\theta(p)} = \overline{\left(\frac{1}{s}\right)}, \tag{11}$$

where $\overline{(1/s)}$ is the average of the reciprocal of the number of symmetries over the star trees with p unlabeled points. Using the results and the notation of Paper III, it is seen that, for large p,

$$\overline{\left(\frac{1}{s}\right)} \sim \frac{ab}{3x_0\delta} \left(\frac{\xi_1}{x_0}\right)^{p-5/2}. \tag{12}$$

One can readily convince one's self that $\xi_0/x_0 < 1$ for any fixed c, and therefore the average number of symmetries grows exponentially with p. However, ξ_0/x_0 becomes closer to 1, and therefore the growth with p will be slower, as c increases. For example, for $c = 0$ (Cayley trees) $\xi_0/x_0 = 0.9196$; for $c = 1$ (mixed Husimi trees) $\xi_0/x_0 = 0.9305$. One sees that while it is not true that for large p the majority of those connected graphs which are star trees built out of stars with no more than c cycles have symmetry number 1, this limit presumably will be reached as c increases.

One of us (G. W. F.) wishes to express his gratitude to Professor Oppenheimer and the Institute for Advanced Study for their hospitality and for a grant in aid which made his stay possible.

[1] G. W. Ford and G. E. Uhlenbeck, "Combinatorial Problems in the Theory of Graphs. I," these PROCEEDINGS, **42**, 122, 1956; G. W. Ford, R. Z. Norman, and G. E. Uhlenbeck, "Combinatorial Problems in the Theory of Graphs. II," these PROCEEDINGS, **42**, 203, 1956; G. W. Ford and G. E. Uhlenbeck, "Combinatorial Problems in the Theory of Graphs. III," these PROCEEDINGS, **42**, 529, 1956. We shall refer to these notes as "Paper I," "Paper II," and "Paper III," respectively, and equations in them will be quoted as "(I, 17)," etc.

[2] R. J. Riddell, "Contributions to the Theory of Condensation" (dissertation, University of Michigan, 1951); R. J. Riddell and G. E. Uhlenbeck, "On the Theory of the Virial Development of the Equation of State of Monoatomic Gases," *J. Chem. Phys.*, **21**, 2056, 1953. The contributions of Polya are given in these references but have not been published by Polya himself.

[3] The number of independent cycles, c, in a graph with p points and k lines is given by the Euler formula $c = k - p + 1$.

[4] Actually it is not sufficient to prove, as in Riddell, *op. cit.*, and Riddell and Uhlenbeck, *op. cit.*, that the correction terms in these expansions are small, since the total number of terms grows with p. One must show that the sum of all correction terms is dominated by the leading term for large p and $k > p \log p$. A rigorous proof is straightforward but is too lengthy to include in this note.

[5] The precise range of k is not known. Polya proves that relation (3) is valid if $|k - {}^1/_4 p(p-1)| < cp$, where c is a positive constant. However, the range is probably larger; in any case, the range certainly covers the majority of graphs.

[6] If there is no root of equation (7) which is less than z_0, then $T(x)$ will still be convergent, but the radius of convergence would be determined by z_0, and no general statement about the analytic behavior of $t(x)$ can be made, since it depends on the behavior of $S(z)$ near z_0.

[7] For the definition of homeomorphic type see D. König, *Theorie der endlichen und unendlichen Graphen* (Leipzig, 1936), p. 5.

[8] We will call a *chain* a line sequence in which the inner points are of order 2, while the end-points are of higher order. These end-points, which are of order 3 or higher, we will call the *principal points* of the homeomorphic type.

[9] The symmetry number of a homeomorphic type we will define as the order of the group of permutations of the principal points which leaves the homeomorphic type invariant.

[10] Compare the discussion in Section 4 of Paper I.

INDICES OF RANK AND OF SINGULARITY ON ABELIAN VARIETIES

BY R. C. GUNNING

DEPARTMENT OF MATHEMATICS, PRINCETON UNIVERSITY

Communicated by S. Bochner, November 2, 1956

The present note is an outline of the formalism of the higher indices of singularity introduced by Comessatti on Abelian varieties[1] and their relation to the rank problem on algebraic varieties.

1. Let M_n be a compact complex manifold; $\Gamma = \sum \Gamma^{(p,q)}$ be the exterior algebra of closed, complex differential forms on M_n, bigraded according to type; $H(M_n, C) = \sum H^d(M_n, C)$ be the complex cohomology algebra of M_n; and $H(M_n, Q) = \sum H^d(M_n, Q)$ be the rational cohomology algebra of M_n considered as a subalgebra of $H(M_n, C)$. The de Rham theory establishes an algebra homomorphism ρ: $\Gamma \to H(M_n, C)$ (onto); the submodule

$$H_k^d = \rho\Big(\sum_{\substack{p+q=d \\ p,\, q \geq k}} \Gamma^{(p,q)}\Big) \cap H^d(M_n, Q)$$

will be called the module of d-dimensional cohomology classes of *semirank* k, and $b_k^d = \dim_Q H_k^d$ the corresponding *index of singularity* of M_n. Let $\theta_1, \ldots, \theta_b \in H_1^2$, $b = b_1^2$, be a Q-basis for H_1^2, and introduce the Q-homomorphisms $L_{i_1 \ldots i_k}: H^{d-k}(M_n, Q) \to H_k^d$ defined by $L_{i_1 \ldots i_k}(\phi) = \theta_{i_1} \cup \ldots \cup \theta_{i_k} \cup \phi$. The Q-submodule $H^*{}_k^d \subset H_k^d$ generated by the images of the homorphisms $L_{i_1 \ldots i_k}$, $1 \leq i_1, \ldots, i_k \leq b_1^2$, will be called the module of d-dimensional cohomology classes of *rank* k, and $c_k^d = \dim_Q H^*{}_k^d$ the corresponding *index of rank* of M_n. If (ϕ, ψ) is a Q-valued inner product on $H^d(M_n, Q)$, and $\Lambda_{i_1 \ldots i_k}: H_k^d \to H^{d-k}(M_n, Q)$ is the algebra homorphism adjoint to $L_{i_1 \ldots i_k}$, then the Q-submodule $H^{**}{}_k^d \subset H_k^d$ which is the intersection of the kernels of the homorphisms $\Lambda_{i_1 \ldots i_k}$ will be called the module of *totally k-effective* d-dimensional cohomology classes. Clearly $H_k^d = H^*{}_k^d \oplus H^{**}{}_k^d$, so that $b_k^d = c_k^d + a_k^d$, where $a_k^d = \dim_Q H^{**}{}_k^d$.

2. These general definitions take particularly simple forms in two cases of note. First, if M_n is a nonsingular algebraic variety with $b_1^2 = 1$, the totally k-effective cohomology classes of dimension $d \leq n$ are dual to the ineffective cycles of classes $\geq k$ in the sense of Lefschetz;[2] the indices introduced above then reduce to well-known invariants. Second, if M_n is a compact complex torus, $H(M_n, C)$ is an exterior algebra isomorphic to the subalgebra Γ_0 of Γ consisting of differential forms with constant coefficients, the isomorphism being determined by the period matrix Ω of the torus, while $H(M_n, Q)$ is isomorphic to the subalgebra of Γ_0 consisting of differential forms with rational periods. The indices of rank and of singularity can then be expressed directly in terms of the period matrix. In particular, let $(\sum \overline{\sum}) = (\Omega/\overline{\Omega})^{-1}$, where $\sum = (\sigma_{ls})$ is a $2n \times n$ complex matrix.

THEOREM 1. *For a complex torus, b_k^d is the dimension of the Q-module of skew-symmetric rational forms $\chi_{j_1 \ldots j_d}$ such that*

$$\sum \chi_{j_1 \ldots j_d} \sigma_{j_1 s_1} \ldots \sigma_{j_{d-k+1} s_{d-k+1}} = 0$$

for all indices $s_1, \ldots, s_{d-k+1}, j_{d-k+2}, \ldots, j_d$.

If M_n is an Abelian variety, it is possible to introduce a Q-valued inner product on $H^d(M_n, Q)$ obtained from an inner product on Γ_0 defined by a Hermitian metric $\sum g_{\mu\bar{\nu}}\, dz^\mu\, d\bar{z}^\nu$ with constant coefficients;[3] let $\chi^{i_1 \ldots i_k}{}_{s_1 \ldots s_{2k}} \in H^{2k}(M_n, Q)$ be rational skew-symmetric forms representing the elements $\theta_{i_1} \cup \ldots \cup \theta_{i_k}$, and define

$$\xi^{i_1 \ldots i_k}{}_{t_1 \ldots t_{2k}} = \sum_{s,\, \mu,\, \nu} g^{\mu_1 \bar{\nu}_1} \ldots g^{\mu_k \bar{\nu}_k} \chi^{i_1 \ldots i_k}{}_{s_1 \ldots s_{2k}} \bar{\sigma}_{s_1 \nu_1} \ldots \bar{\sigma}_{s_{k} \nu_k}$$

$$\cdot \sigma_{s_{k+1} \mu_{k+1}} \ldots \sigma_{s_{2k} \mu_{2k}} \sigma_{t_1 \mu_1} \ldots \sigma_{t_k \mu_k} \bar{\sigma}_{t_{k+1} \nu_{k+1}} \ldots \bar{\sigma}_{s_{2k} \nu_{2k}}.$$

THEOREM 2. *For an Abelian variety, a_k^d is the dimension of the Q-module of skew-symmetric rational forms $x_{j_1 \ldots j_d}$ such that*

$$\sum_{t_1 \ldots t_{2k}} \xi^{i_1 \ldots i_k}{}_{t_1 \ldots t_{2k}} \chi_{t_1 \ldots t_{2k} t_{2k+1} \ldots t_d} = 0$$

for all indices $i_1, \ldots i_k, t_{2k+1}, \ldots, t_d$.

It should be noted that on Abelian varieties, b_1^d is the d-dimensional index of singularity of the Riemann matrix Ω in the sense of Comessatti.[1]

3. In conclusion, the geometrical interpretation of the indices of rank on an algebraic variety should be observed.

THEOREM 3. *On a nonsingular algebraic variety, c_k^d is the dimension of the Q-submodule of $H^d(M_n, Q)$ generated by cohomology classes dual to $(2n - d)$-dimensional rational cycles formed by intersecting $(2n + 2k - d)$-dimensional rational cycles with $(n - k)$-dimensional algebraic subvarieties of M_n.*

Thus the index c_{n-k}^{2n-d} in a sense gives the dimension of the set of d-dimensional cycles lying on k-dimensional algebraic subvarieties of M_n. The proof of this result follows from a theorem of Lefschetz,[2] which asserts that a $(2n - 2)$-dimensional cycle is represented by a divisor (effective or not) if and only if its dual cohomology class is represented by a differential form of type (1, 1), and from a theorem of Severi,[4] which asserts that an irreducible algebraic subvariety $V_k \subset M_n$ is the complete intersection of $r - k$ divisors (effective or not) on M_n.

[1] A. Comessatti, "Sugl'indici di singolarità a più dimensioni delle varietà abeliane," *Rend. Seminar. mat. Univ. Padova*, Vol. 5, 1934.

[2] W. V. D. Hodge, *Harmonic Integrals* (Cambridge, 1952).

[3] This corresponds to the Riemann conditions on Ω.

[4] F. Severi, *Serie, sistemi d'equivalenza e corrispondenza algebriche sulla varietà algebriche* (Rome, 1942).

ON DEHN'S LEMMA AND THE ASPHERICITY OF KNOTS

BY C. D. PAPAKYRIAKOPOULOS

INSTITUTE FOR ADVANCED STUDY

Communicated by Deane Montgomery, November 15, 1956

Everything in this note will be considered from the *semilinear* point of view; i.e., any 3-manifold will be considered with a fixed triangulation (this is permissible according to Moise's work[1]), any curve or line will be considered as polygonal, any surface as polyhedral, and so on.

The following theorem was first considered by Dehn,[2] but it was pointed out by Kneser[3] that Dehn's proof contains a gap.

DEHN'S LEMMA. *Let M be a 3-manifold, compact or not, with boundary which may be empty, and in M let D be a 2-cell with self-intersections (singularities), having as boundary the simple closed polygonal curve C and such that there exists a closed neighborhood of C in D which is an annulus (i.e., no point of C is singular). Then there exists a 2-cell D_0 with boundary C, semilinearly imbedded in M.*

Johansson[4] proved that, if Dehn's lemma holds for all orientable 3-manifolds, it also holds for all nonorientable ones. We prove that Dehn's lemma holds for all *orientable* 3-manifolds, and by a modification of our method we also prove the following theorem.

SPHERE THEOREM. *Let M be an orientable 3-manifold, compact or not, with boundary which may be empty, such that $\pi_2(M) \neq 0$, and which can be topologically imbedded in a 3-manifold N, having the following property: The first homology group of any non-trivial (but not necessarily proper) subgroup of $\pi_1(N)$, has an element of infinite order (note in particular that this holds if $\pi_1(N) = 1$). Then there exists a 2-sphere S semilinearly imbedded in M, such that S is not homotopic to zero in M.*

From these two theorems follow others, which may be stated easily if we introduce the following notions: Let F be a nonempty proper closed subset in S^3. We say that F is *geometrically splittable* if there is a 2-sphere $S^2 \subset S^3 - F$ such that both components of $S^3 - S^2$ contain points of F. We say that F is *algebraically splittable* if $\pi_1(S^3 - F)$ is the free product of two groups, each of which is nontrivial.

THEOREM 1. *Let U be a nonempty proper open connected subset of the 3-sphere S^3. Then U is aspherical if and only if $S^3 - U$ is not geometrically splittable.*

The above theorem provides us with a solution of a problem of Whitehead.[5] From Theorem 1 follows easily the following:

COROLLARY 1. *If F is a nonempty proper closed connected subset of S^3, then each component of $S^3 - F$ is aspherical.*

The above corollary provides us with a solution of a problem of Eilenberg.[6] An immediate consequence of Corollary 1 is the following:

COROLLARY 2. *If F is a connected graph or knot, then $S^3 - F$ is aspherical.*

The following theorem solves completely a problem initiated by Higman.[7]

THEOREM 2. *Let K be a link in S^3. The following three statements are equivalent:*

(*i*) *$S^3 - K$ is not aspherical.*
(*ii*) *K is geometrically splittable.*
(*iii*) *K is algebraically splittable.*

Let K be a *knot* in S^3. According to Dehn (*op. cit.*, Satz 2, p. 158), Specker,[8] Papakyriakopoulos,[9] and the above Corollary 2, the following theorem holds:

THEOREM 3. (*i*) *K is unknotted if and only if $\pi_1(S^3 - K)$ is free cyclic.* (*ii*) *The number of ends of $\pi_1(S^3 - K)$ is either 1 or 2.* (*iii*) *$\pi_1(S^3 - K)$ has 1 end if K is knotted and 2 ends if K is unknotted.*

The following statement is known as *Hopf's conjecture*.

THEOREM 4. *If U is an open connected subset of the 3-sphere, then $\pi_1(U)$ has no element of finite order.*

The proof is based on the sphere theorem and the following simple consequence of a theorem due to P. A. Smith:[10] The fundamental group of an aspherical polyhedron (of finite dimension) has no element of finite order.

A Sketch of the Proof of Dehn's Lemma.—By a *Dehn disk* we mean a 2-cell, which may have singularities, but not on its boundary. We suppose that M is *orientable* and that D has no *branch points* (see Whitehead[11]). Let G be the *inverse image* of D under f, where G is a 2-cell. There is a finite set J of *triples* (J', J'', ψ), where J', J'' are closed curves on G, called the *J-curves*, and ψ: $J' \to J''$ is a "nice map" such that $f(r) = f\psi(r)$, for any point $r \in J'$, i.e., $f(J') = f(J'')$ is a *double line*[11] of D. We emphasize that, according to Johansson,[12] $J' \neq J''$ because M is orientable. So we have a map f: $(G, J) \to D$ called a *realized diagram*. We denote by $t(D)$ and $d(D)$ the number of *triple points*[11] and double lines of D, respectively. The couple $(t(D), d(D))$ is called the *complexity* of D.

Let $V \subset M$ be a 3-manifold with boundary such that int V is a neighborhood of $D - C$, $C \subset$ bdry V, and V is very small and "very nice," so that each face of bdry V is "parallel" to a face of D. Such a V is called a *prismatic neighborhood* of D in M. We observe that D is a deformation retract of V. Let p_*: $M_* \to V$ be the universal covering of V, let D_* be a Dehn disk covering D just once, and let

$q_* = p_*|D_*$. Finally, let V_* be a prismatic neighborhood of D_* in M_*. So we have the following diagram:

$$\begin{array}{ccccccc} M_* & \supset & V_* & \supset & D_* & \xleftarrow{f_*} & (G, J_*) \\ \downarrow{\scriptstyle p_*} & & & & \downarrow{\scriptstyle q_*} & & \\ M & \supset & V & \supset & D & \xleftarrow{f} & (G, J) \end{array}$$

called an *elementary tower* over $D \subset V \subset M$, where f_* is the lifted map, and f_*: $(G, J_*) \to D_*$ is a realized diagram.

If V is *not* simply connected, then there is a covering translation τ of p_*: $M_* \to V$, such that $D_* \cap \tau^{-1}(D_*)$ is not empty, and so it consists of a finite number of closed curves T_{*i}, $i = s(\tau, 1), \ldots, s(\tau, d(\tau))$. These curves for all possible τ's are called the T_*-*curves*. Then

$$d(D) = d(D_*) + \frac{1}{2}\sum d(\tau),$$

where the sum runs over all covering translations $\tau \neq 1$ (note that if $D_* \cap \tau^{-1}(D_*) = \varnothing$, then naturally $d(\tau) = 0$). Hence $d(D_*) < d(D)$. We emphasize that, in the special case where D_* is a 2-*cell*, q_*: $(D_*, T_*) \to D$ is a *realized diagram*, where T_* is now the set of all triples (T', T'', τ) such that $T' \in D_* \cap \tau^{-1}(D_*)$ and $T'' = \tau(T') \in \tau(D_*) \cap D_*$.

If V is *simply connected*, then each of the components of bdry V is a 2-*sphere*, according to a result due to Seifert.[31] In this case $d(D_*) = d(D)$.

The diagram

$$\begin{array}{cccccccc} & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & & & & & & \\ & M_m & \supset & V_m & \supset & D_m & \xleftarrow{f_m} & (G, J_m) \\ (m) & \downarrow{\scriptstyle p_m} & & & & \downarrow{\scriptstyle q_m} & & \\ & M_{m-1} & \supset & V_{m-1} & \supset & D_{m-1} & \xleftarrow{f_{m-1}} & (G, J_{m-1}) \\ & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & & & & & & \\ & M_1 & \supset & V_1 & \supset & D_1 & \xleftarrow{f_1} & (G, J_1) \\ (1) & \downarrow{\scriptstyle p_1} & & & & \downarrow{\scriptstyle q_1} & & \\ & M & \supset & V & \supset & D & \xleftarrow{f} & (G, J) \end{array}$$

is called a *tower* over $D \subset M$ and is defined as follows: The diagram (m) is an elementary tower over $D_{m-1} \subset V_{m-1} \subset M_{m-1}$, for $m = 1, 2, \ldots$, where $M_0 = M$, $V_0 = V$, $D_0 = D$, $f_0 = f$, $J_0 = J$. Then

$$d(D_0) \geqq d(D_1) \geqq d(D_2) \geqq \ldots \geqq 0,$$

and there exists a number $n \geqq 0$, such that $d(D_i) > d(D_{i+1})$, for $i < n$, and $d(D_j) = d(D_{j+1})$ for $j \geqq n$, i.e., V_n is *simply connected* but V_i is *not*. This number n is called

the *height* of the tower. The following two cases are possible: (1) $d(D_n) > 0$, i.e., D_n is not a 2-cell; (2) $d(D_n) = 0$, i.e., D_n is a 2-cell.

In case (1) V_n is simply connected, and so bdry V_n is composed of 2-spheres, according to the result of Seifert mentioned above. Let D_n' be one of the 2-cells bounded by C_n, on the component of bdry V_n containing the boundary C_n of D_n. Then $D' = p_1 \ldots p_n(D_n') \subset M$ is a Dehn disk with boundary C and "roughly speaking" having the following property: Either $t(D') < t(D)$, or $t(D') = t(D)$ and $d(D') < d(D)$, i.e., we may say that *complexity of* $D' <$ *complexity of* D.

In case (2) we prove that there exists a triple (J', J'', ψ) of f: $(G, J) \to D$ such that J' and J'' are *disjoint simple* closed curves. The proof of this is rather algebraic and makes use of the fact that q_n: $(D_n, T_n) \to D_{n-1}$ is a realized diagram, as we have observed above, and that each element of T_n is a triple (T', T'', τ), where τ is a covering translation of p_n: $M_n \to V_{n-1}$. Then by a *cut* ("Umschaltung";[14] note that this is the only case in which we can apply Dehn's process without any danger [see Dehn, *op. cit.* p. 150, B, and Kneser, *op. cit.*, p. 260]) of D along $J = f(J') = f(J'')$ we obtain a new Dehn disk $D' \subset M$, with boundary C and such that either $t(D') < t(D)$ or $t(D') = t(D)$ and $d(D') < d(D)$, i.e., we may say that *complexity of* $D' <$ *complexity of* D.

In the same way we obtain from D' a new Dehn disk $D'' \subset M$ with boundary C such that complexity of $D'' <$ complexity of D', and so on. Thus, after a finite number of repetitions of the above process, we finally obtain a Dehn disk in M with boundary C and complexity (0, 0), i.e., we obtain a 2-cell in M with boundary C.

As far as the proof of the sphere theorem is concerned, we restrict ourself to the remark that the method of proof makes use of the above process, standard Hurewicz theorems, and the Poincaré duality theorem.

[1] E. E. Moise, *Ann. Math.*, **56**, 96–114, 1952, and **59**, 159–170, 1954.
[2] M. Dehn, *Math. Ann.*, **69**, 147, 1910.
[3] H. Kneser, *Jahresber. Deut. math. Verein.*, **38**, 260, 1929.
[4] I. Johansson, *Math. Ann.*, **115**, 658–669, 1938, sec. 2.
[5] J. H. C. Whitehead, *Fundamenta math.*, **32**, 161, 1939.
[6] S. Eilenberg, *Fundamenta math.*, **28**, 241, 1937.
[7] G. Higman, *Quart. J. Math.* (Oxford), **29**, 117–122, 1948.
[8] E. Specker, *Comm. Math. Helv.*, **23**, 329, 1949.
[9] C. D. Papakyriakopoulos, *Ann. Math.*, **62**, 298, 1955, Corollary 3.
[10] W. Hurewicz, *Proc. Acad. Sci. Amsterdam*, **39**, 216, 1936.
[11] J. H. C. Whitehead, *J. London Math. Soc.*, **12**, 63–71, 1937, first footnote on p. 66.
[12] I. Johansson, *Math. Ann.*, **110**, 319, 1935, 2. Satz.
[13] H. Seifert and W. Threlfall, *Lehrbuch der Topologie*, Satz IV, 223, 1934.
[14] J. H. C. Whitehead, *J. London Math. Soc.*, **12**, 63–71, 1937, second footnote on p. 66

FUNCTION RINGS ON THE CIRCLE

By J. Wermer

Brown University and Institute for Advanced Study

Communicated by S. Bochner, October 31, 1956

1. *Introduction.*—Let C be the normed ring of all continuous complex-valued functions on the unit circle S_0: $|u| = 1$. If χ is a homeomorphism of S_0 onto a space S, we let $\tilde{\chi}$ denote the induced isomorphism on C, defined for g in C by $\tilde{\chi}g(s) = g(\chi^{-1}s)$ for all s in S.

Let A be a subring of C containing all constants and separating points on S_0. Denote by $\bar{A}$ the closure of A. When is $\bar{A} \neq C$?

We say that A is of *type R* if there exists a homeomorphism χ of S_0 on a simple closed curve $\tilde{\gamma}$ which lies on a Riemann surface $\tilde{\mathfrak{F}}$ such that

(I) $\tilde{\gamma}$ bounds a region $\tilde{\mathfrak{D}}$ on $\tilde{\mathfrak{F}}$ with $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$ compact.

(II) For all g in A, $\tilde{\chi}\, g$ can be extended to $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$ to be continuous there and analytic on $\tilde{\mathfrak{D}}$.

If A is of type R, then $\bar{A} \neq C$. For the maximum principle on $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$ implies that $\bar{A}$ contains only boundary functions of functions analytic on $\tilde{\mathfrak{D}}$ and so contains, e.g., no nonconstant real functions.

In this paper we prove the converse for a class of rings. Let φ and f be a pair of functions in C separating points on S_0. Let $A = A(\varphi, f)$ be the ring of all polynomials in φ and f, including constants. In an earlier paper[1] the author dealt with the case when $\varphi(u) = u$ or $\varphi(u) = u^2$ and f is arbitrary. Here we assume the following:

a) φ and f are analytic in an annulus containing S_0.

b) $\varphi' \neq 0$ on S_0.

THEOREM: *Under hypotheses a, and b,* $\bar{A} \neq C$ *only if A is of type R.*

Note: Without loss of generality we may suppose that φ takes only a finite number of values more than once on S_0.

2. *Preliminaries.*—Denote by γ the image of S_0 under φ. Then γ is a simple closed analytic curve with a finite set M of multiple points. Fix u_0 in S_0, and consider the function element $f(\varphi^{-1})$ at $\varphi(u_0)$. Let $\tilde{\mathfrak{F}}$ be the totality of regular and algebraic places obtained by continuation from this function element. $\tilde{\mathfrak{F}}$ then carries the structure of a Riemann surface.

On any collection of places $\tilde{S}$ there are two distinguished functions z and w, which we call respectively the *projection* function and the *canonical* function on $\tilde{S}$, such that z assigns to each place the complex number over which it lies and w assigns to each place $\tilde{p}$ the value $h(z)$, where h is the function element of $\tilde{p}$ and $z = z(\tilde{p})$.

We define a map χ of S_0 into $\tilde{\mathfrak{F}}$ as follows: for each u in S_0, $\chi(u)$ is the place of $\tilde{\mathfrak{F}}$ obtained by continuing $f(\varphi^{-1})$ along the arc on γ onto which φ maps the positively oriented arc on S_0 joining u_0 to u. Let $\tilde{\gamma}$ be the image of S_0 under χ. Then, since φ and f separate points on S_0, $\tilde{\gamma}$ is a simple closed curve on $\tilde{\mathfrak{F}}$ and χ is a homeomorphism of S_0 on $\tilde{\gamma}$.

To prove our theorem, making use of the objects $\tilde{\mathfrak{F}}$, $\tilde{\gamma}$, χ just introduced, we need to do the following: we must find a region $\tilde{\mathfrak{D}}$ on $\tilde{\mathfrak{F}}$ bounded by $\tilde{\gamma}$ such that $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$

is compact, and we must show that $\tilde{\chi}\varphi$ and $\tilde{\chi}f$ may be analytically extended from $\tilde{\gamma}$ to $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$. For, since every g in A is a polynomial in φ and f, it would follow from this that condition II is satisfied for all g in A.

Finally, we observe that $\tilde{\chi}\varphi$ and $\tilde{\chi}f$ agree on $\tilde{\gamma}$ respectively with the projection function and the canonical function on $\tilde{\mathfrak{F}}$. It will suffice, then, to show that these two functions are analytic on $\tilde{\mathfrak{D}}$.

3. *Regular Surfaces.*—For each λ in $\gamma - M$ and each g in A, we set $g^*(\lambda) = g(\varphi^{-1}(\lambda))$. We write Ω for the complement of γ. Since $\bar{A} \neq C$, we can find a measure $d\mu$ on $\gamma - M$ with $d\mu$ orthogonal to g^* for each g in A, and $d\mu \not\equiv 0$. Now fix a component W of Ω. For each g in A we define a function ϕ on W as follows:

$$\phi(W, g, z) = \frac{1}{2\pi i}\int_{\gamma}\frac{g^*(\lambda)d\mu(\lambda)}{\lambda - z}, \qquad z \in W. \tag{1}$$

Clearly ϕ is analytic in $\bar{W}$.

Let $\bar{W}$ be an m-sheeted Riemann surface over W defined by an equation

$$\sum_{i=0}^{m} c_i(z)w^i = 0, \tag{2}$$

where the c_i are meromorphic on W and $c_m \equiv 1$. Assume the discriminant of (2) $\not\equiv 0$.

Definition: $\bar{W}$ is *regular* over W if there exists a meromorphic function k on $\bar{W}$, $k \not\equiv 0$, and a homomorphism L of $\bar{A}$ into functions meromorphic on $\bar{W}$ such that Lf is the canonical function on $\bar{W}$ and such that

$$\sum_{i=1}^{m} Lg(\tilde{p}_i)\, k(\tilde{p}_i) = \phi(W, g, z) \tag{3}$$

for all g in A and all z in W which are not branch points, where $\tilde{p}_1, \ldots, \tilde{p}_m$ are the places on $\bar{W}$ lying over z.

It is not clear a priori that a regular surface exists over a given W. However, we have

LEMMA 1. *Over a given W there exists at most one regular surface.*

LEMMA 2. *Let the regular surface over a certain W exist and be defined by equation* (2). *Then the coefficients c_i in equation* (2) *are analytic in W and at each boundary point of W.*

COROLLARY. *The places defined by equation* (2) *over the closure of W form a compact set of places $(\bar{W})_c$ containing $\bar{W}$. Also, the canonical function on $\bar{W}$ is analytic everywhere on $(\bar{W})_c$.*

4. *Proof of the Theorem.*—We have

LEMMA 3. *For a certain component W_0 of Ω, the regular surface $\bar{W}_0$ over W_0 exists and is one-sheeted. Also $\bar{W}_0$, considered as a collection of places, is a region on $\tilde{\mathfrak{F}}$.*

We now define a region $\tilde{\mathfrak{D}}$ on $\tilde{\mathfrak{F}}$ to be that component of the complement of $\tilde{\gamma}$ on $\tilde{\mathfrak{F}}$ which contains $\bar{W}_0$.

LEMMA 4: *Let $\tilde{p} \in \tilde{\mathfrak{D}}$. Then there exists a bounded component W of Ω such that the regular surface $\bar{W}$ over W exists and such that $\tilde{p}$ lies in $(\bar{W})_c$. Further, the canonical function on $\tilde{\mathfrak{F}}$ agrees at $\tilde{p}$ with the canonical function on $\bar{W}$.*

It follows that $\tilde{\mathfrak{D}}$ is contained in the union of a finite number of the $(\bar{W})_c$. Since each $(\bar{W})_c$ is compact, $\tilde{\mathfrak{D}}$ has compact closure in $\tilde{\mathfrak{F}}$, and $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$ is compact. Since

for each W the canonical function on $\bar{W}$ is analytic on $(\bar{W})_e$, it follows that the canonical function on $\tilde{\mathfrak{F}}$ is analytic everywhere on $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$. Clearly also the projection function on $\tilde{\mathfrak{F}}$ is analytic on $\tilde{\mathfrak{D}} \cup \tilde{\gamma}$. By the discussion in Section 2, the theorem is thus proved.

[1] J. Wermer, "Algebras with Two Generators," *Am. J. Math.*, **76**, No. 4, 853–859, 1954.

---

*INFLUENCE OF THE TYPE OF INSEMINATED FEMALE ON THE LETHAL FREQUENCY IN THE X CHROMOSOME FROM THE MALE**

BY PHILIP E. HILDRETH† AND GWENETH L. CARSON

DEPARTMENT OF ZOOLOGY AND RADIATION LABORATORY, UNIVERSITY OF CALIFORNIA, BERKELEY

Communicated by Curt Stern, November 2, 1956

Experiments dealing with spontaneous lethal mutation rates in the X chromosome of *Drosophila melanogaster* males have been carried out in this laboratory since 1952. For reasons peculiar to the various experiments, females of different genotypes were used. In checking the lethal frequencies in the X chromosomes from wild-type Samarkand males when these different females were used, we noticed that the frequencies varied considerably, depending on which type of female was mated with the wild-type males.

In our early experiments we used three different X chromosomes bearing inversions, which were kindly supplied to us by Dr. H. J. Muller, of the University of Indiana. The two females used in these experiments had inversions in their X chromosomes and were homozygous Basc, formerly called *Muller-5*, "sc^{s1} *B InS* w^a sc^s," and heterozygous Bv "*y* sc^{s1} *B In49 v* w^a sc^s/y *v f* · *YL*." The males used were Basc and Binscy "*y* sc^{s1} *B In49* sc^s."

Virgin Basc or Bv females were collected and mated with Samarkand males. The parent flies were cultured in two ways; for Sets 8–20, 47–55, 60, 60A, 61, and 61A the flies were mass-mated in half-pint culture bottles containing the usual corn meal, molasses, dried yeast, and agar medium. For Sets 33–35 one male was mated with from four to ten females in a "creamer" (1.5-ounce glass vessel) containing the same medium as above, but each female was transferred individually to a separate creamer the next day. On the fourteenth day (Sets 33–35, 60, and 61) or the fifteenth day (Sets 8–20 and 47–55), the offspring were counted and the B/+ daughters bearing the X chromosome from the P male were mated individually with Basc males (Sets 8–33) or Binscy males (Sets 47–61) in creamers. On the fifteenth day after these matings the flies in the creamers were examined in order to determine whether a lethal had originated in the X chromosome contributed by the original P male. If no lethal was present, the F_2 culture usually contained B/B and B/+ females and + and B males. If such a lethal was present, the culture contained no + males but had the other three classes. For all cultures having at least one + male the Samarkand X chromosomes were designated "not lethal" and the cultures were discarded. For those having 12 or fewer flies but no + males the Samarkand X chromosomes were called "doubtful," and new cultures

were started with the F_1 B/+ female and (or) her daughters. If the culture had 12 or more flies but no + males, the Samarkand X chromosome was provisionally designated "lethal-containing," and one to three new creamers were started, with from one to three B/+ daughters and five Basc males per creamer. The offspring from these new F_2 cultures were checked, and if at least 50 flies and no + males from all F_3 creamers derived from the P male were found, the provisional lethal was considered a definite lethal. If there were fewer than 50 flies and no + males, new cultures were again started whenever possible. The X chromosomes of some F_1 females could not ever be classified as lethal or nonlethal, either because the F_1 females were sterile (steriles) or because their progeny included no + male and fewer than 50 flies in any single generation (doubtfuls). The data concerning these experiments are presented in Table 1.

The total frequency of lethals in the X chromosomes of the original P males derived through the Basc lines was 8 in 6,784 tested chromosomes (0.118 per cent), and the frequency of lethals derived through the Bv lines was 20 in 2,180 tested chromosomes (0.917 per cent). Percentages of lethals were based on the numbers of X chromosomes that could be classified as lethal or nonlethal. Twenty-one lethals were found in the Bv sets, but two of these, (from one female in Set 61) appeared to have identical loci and are assumed to have arisen as a single event early in the germ line of the male. They are counted as one lethal for the calculations above. The frequency of the lethals from the Bv lines is 7.8 times that from the Basc lines. The χ^2 test comparing the values for Bv and Basc gave a χ^2 value of 33.86, with a P value of much less than 0.001. The difference must be attributed to some factor other than chance.

TABLE 1

FREQUENCY OF LETHALS RECOVERED FROM THE X CHROMOSOMES OF MALES WHICH HAD INSEMINATED BASC OR BV FEMALES

SETS	AGE IN HOURS DURING MATINGS: MALES	AGE IN HOURS DURING MATINGS: FEMALES	HOURS MATED	TOTAL* CHROMOSOMES	LETHALS	PER CENT LETHAL	χ^2	P VALUE	YEAR
Basc females:									
SC 8–20	0.05–30.1	0.12–53.9	±24	3,849	3	0.078	1.21	0.28	1952
SC 33–35	0.08–29.1	23.2–53.4	±24	2,935	5	0.170			1955
				6,784	8	0.118			
Bv females:									
SC 47–55	0.03–77.1	0.07–76.7	±48	699	7	1.001	2.15	0.35	1954
SC 60–61	0.07–29.0	0.05–51.9	±24	1,245	9	0.723			1955
SC 60A–61A	23.6–52.8	0.07–47.6	±24	236	4	1.695			1955
				2,180	20	0.917			

* Excluding doubtfuls and steriles.

The early experiments with Bv and Basc females for which data are given in Table 1 were carried out independently of each other. In order to test the reality of a "spouse effect" on rates of recovered lethals, a more carefully controlled experiment was done in which single males were given the opportunity to mate with both types of females. Furthermore, records were kept of the ages of both males and females, and, in addition, it was noted which of the two females available to the male was inseminated first. The latter information is important for the following reason. If there is a true difference in the rate of lethals recovered, the event

responsible for the difference may occur either before or after insemination of the females. If it occurs before insemination, then the different samples of sperm transferred in mating might be responsible for the difference in lethal frequency, as it is known that the lethal frequency varies with the age of the male.[1] In order to determine whether sperm in the first and second inseminations differed in mutation rate, a comparison was necessary between the rates of lethals recovered when the Basc females were mated first and second and when the Bv females were mated first and second. If the male preferentially mated first with one type of female, this could possibly have accounted for the difference between the lethal frequencies derived through the Basc and Bv lines.

Method.—Bottles of the three stocks—wild-type Samarkand, Bv, and Basc—were emptied in the morning. Within two to four hours afterward, virgin Bv and Basc females and + males were collected. The flies were then stored, one Bv and one Basc female together in one vial and each male in a separate vial, until the next afternoon. The vials were approximately 19 mm. in diameter and 65 mm. deep and contained a medium of molasses, water, agar, and dried yeast. Early in the afternoon of the next day, approximately 50 males were shaken, individually without etherization, into separate vials each containing the two females, and the time was recorded. To observe the mating behavior of the flies, the vials were laid on their sides, alternating with empty vials, in three or four rows on a tin tray approximately 45 × 30 × 2.2 cm. deep. The bottom of the tray was lined with white paper. From the time the first male was placed with its two females, each vial was checked at 5-minute intervals. Records were kept of whether the male was mating with the Basc (apricot eye color) or Bv (vermilion eye color) female or not mating with either. No male used for progeny tests was observed mating more than once with each kind of female. When a male had completed mating with both females, the three flies were etherized, each female was placed in a separate empty vial, and the male was discarded. When the two females had awakened, each was shaken into a separate half-pint culture bottle labeled with the type of female, the number of the male, and the time that the flies were placed in the bottles. Forty-eight hours after being placed in the bottles the females were removed and discarded. Twelve days after this the bottles were examined, and only those pairs of bottles were used in which both the Bv and Basc females produced F_1 B/+ daughters from the same male. The B/+ daughters were stored with at least twice as many Basc males in creamers; up to 25 females were stored in one creamer. After two days' storage the females were placed individually in creamers with two Basc males, and the creamer was labeled with the type of female and the number of the P male. Fifteen days later these cultures were examined to determine whether they contained lethal X chromosomes from the original wild-type male. The criteria and methods used in the experiments reported earlier were also used in these tests. While the flies were in the mating vials, the temperature varied from 21.6° to 27.0° C. because of the variation in room temperature (except for Set S46A, which had a maximum temperature of 33.7° C. and considerable loss of fertility). At all other times the flies were kept at about 26° C. in incubators while being stored or cultured.

Results.—Twenty-three separate experiments were carried out, and, of the males tested, 191 had mated first with a Basc female and 134 had mated first with a Bv female. The difference is significant ($\chi^2 = 10.00$; $P = 0.0028$). The minimum

and maximum ages of the males when mated with the Basc females ranged from 23.7 to 37.0 hours, as compared with a range of from 23.7 to 36.9 hours for the matings with the Bv females. Thus there was little difference in the range of ages of the males when mated with the two types of females. The Basc females were from 24.9 to 36.9 hours old when mated, and the Bv females from 23.5 to 37.2 hours old; therefore, there was little difference in the range of ages of the Basc and Bv females when they were mated. Table 2 summarizes the results of the experiments, giving the lethal frequencies in the X chromosome from the male when mated first with the Basc female and second with the Bv female and also when the mating order was reversed. The values in the table represent the number of X chromosomes tested from the Samarkand males.

TABLE 2

FREQUENCY OF LETHALS RECOVERED THROUGH BASC AND BV FEMALES

	Basc Mated First			Bv Mated First			Combined Data (Total* Per Cent Lethal)
	Total*	Lethal	Per Cent Lethal	Total*	Lethal	Per Cent Lethal	
Basc	7,562	17	0.225	3,330	7	0.210	0.220
Bv	2,727	12	0.440	2,659	20	0.752	0.594

* Excluding doubtfuls and steriles.

The χ^2 test of the difference in lethal frequency between the Basc females when mated first and second gives a value of 0.023, which has a P value of 0.89, and the test between the Bv females when mated first and second gives a value of 2.22, with a P value of 0.15. Thus the order of mating makes little difference as far as the lethal frequencies are concerned. In order to determine whether tests using Basc females gave different mutation rates than tests using Bv females, the following χ^2 tests were made. In the sets in which the Basc female mated first and the Bv female second, the χ^2 value is 3.30 and P is 0.074. In the sets where the mating order was reversed, χ^2 is 9.68 and P is 0.0035. The χ^2 test between the total Basc and Bv sets, counting all lethals, gives a value of 14.69, with a P value of less than 0.001.

In the sets in which the Basc females mated first, there was one Basc female from which eight lethals were recovered. If this cluster of lethals is considered a single lethal instead of eight, then the χ^2 value, testing the significance between the frequency of recovered lethals from the two types of females when Basc mated first, is 8.89, with a P value of 0.0052. The χ^2 value for the difference between the Basc females when mated first and second if a cluster is counted as a single lethal is changed from the 0.023 given above to 0.90, and P is 0.36. If the cluster is considered a single lethal, the χ^2 value for all Basc and Bv sets is 23.02, with a P value of less than 0.001. If the lethals in the cluster are counted as eight, the frequency of lethals recovered from the Bv females (0.594 per cent) is 2.70 times that recovered from the Basc females (0.220 per cent), while if the cluster is counted as a single lethal, the Bv frequency (0.594 per cent) is 3.81 times that of the Basc (0.156 per cent). In either case the difference between tests with Bv and Basc females is significant, with a P value of less than 0.001.

To analyze these data further, we used the nonlethal and lethal chromosomes

from only those pairs of cultures in which at least one lethal appeared in either member of the pair.[2] This method eliminated the necessity of assuming that the lethal frequency was the same from one male to another. A value Z was computed such that

$$Z = \frac{\sum_i X_i - \sum_i \mu_i}{\sqrt{\sum_i \sigma_i^2}}, \qquad (1)$$

where

$$\sigma_i^2 = \frac{n_i N_{1i} N_{2i} (N_i - n_i)}{N_i^2 (N_i - 1)}, \quad \text{except where } n_i = 1; \text{ then } \sigma_i^2 = q_i(1 - q_i),$$

i = any pair where at least one lethal occurs,
N_{1i} = number of nonlethal plus lethal offspring from Bv,
N_{2i} = number of nonlethal plus lethal offspring from Basc,
n_i = number of lethal offspring from Bv plus Basc,
X_i = number of lethal offspring from Bv,
$N_i = N_{1i} + N_{2i}$,
$q_i = N_{1i}/N_i$, and
$\mu_i = n_i q_i$.

In the sets where Basc females mated first, 9 males produced lethals in the Basc cultures, 10 produced lethals in the Bv cultures, and 1 produced lethals in both cultures. In the Bv cultures there were 12 lethals in 337 chromosomes tested, and in the Basc cultures there were 17 lethals in 719 chromosomes. The Z value for these data is 1.97, with a P value of 0.024 for the probability of obtaining by chance a difference between the Bv and Basc cultures as large as or larger than that found if the frequencies were actually equal. Data are given in Table 3. In the sets in which Bv females mated first, 18 males in the Bv cultures produced 20 lethals in 560 chromosomes tested, and 6 males in the Basc cultures produced 7 lethals in 566 chromosomes tested. The Z value for these data is 2.40, with a P value of 0.0082. Data are given in Table 3.

The statistical analysis included all lethals from a cluster that might have resulted from the multiplication, during mitosis and meiosis, of a single lethal that arose early in the germ line of the original male. This inclusion of all recovered lethals is necessary if one assumes that the difference in the frequency of male X-chromosome lethals recovered from the Basc and Bv females is caused by some event that occurred after insemination of these females. However, whenever more than one lethal came from a single male, they were localized by means of crossover tests to determine whether they were identical or had arisen as separate events. There was one male whose offspring from the Bv female gave one lethal and whose offspring from the Basc female gave eight lethals. The crossover tests showed these to be probably allelic lethals near the yellow locus. Three other males had more than one lethal-bearing daughter. The Bv mates of two of the males had two and three daughters, respectively, that carried nonallelic lethals. The Basc mate of the third male had two daughters that carried nonallelic lethals.

Discussion.—There are several factors that could have influenced the frequency of lethals. If, for example, the Basc cultures had been more crowded than the Bv cultures, more of the heterozygotes for lethals might have died before developing into adults in the Basc cultures. Lethal heterozygotes have been shown to be, on the average, less viable than similar heterozygotes that did not carry lethals (Stern, Carson, Kinst, Novitski, and Uphoff[3]). However, the sizes of the corresponding populations in the Basc and Bv crosses were similar. The average numbers of offspring for Bv and Basc females mated first and second were as follows: Basc first, 83.3, Bv first, 87.4; Basc second, 53.9, Bv second, 58.9. Crowding appears to have had no effect on lethal frequency as far as the increase of Bv over Basc tests is concerned.

TABLE 3

Z TEST

	Basc Females Mated First					Bv Females Mated First			
	Basc Females		Bv Females			Basc Females		Bv Females	
Set	Non-lethal	Lethal	Non-lethal	Lethal	Set	Non-lethal	Lethal	Non-lethal	Lethal
S42A	16	1	3	0	S42B	14	0	28	1
S42C	68	1	11	0	S43A	19	1	26	0
	25	0	34	1		9	0	25	1
S43A	48	0	8	1	S44C	16	0	32	1
	44	1	3	0		49	1	39	0
S43B	39	0	11	1		27	1	21	0
	34	1	23	0	S45A	17	0	19	1
S44B	36	1	32	0	S45B	21	0	26	1
	42	1	31	0	S46A	14	0	25	1
S44C	37	1	30	0	S46B	25	0	13	1
S45B	43	8	6	1		6	0	15	1
S45C	24	0	11	1	S46C	17	1	29	0
S47A	25	0	21	1		11	0	11	1
S47B	30	0	21	1	S47A	28	0	28	1
	51	1	5	0		8	0	22	3
S48A	9	0	13	1	S47B	1	0	18	1
S48C	13	0	8	1		37	0	29	1
S49B	50	0	15	2	S48B	36	0	3	1
	30	1	23	0		28	0	4	1
S49C	38	0	16	1	S48C	34	2	16	0
					S49A	53	1	6	0
						15	0	37	1
					S49B	54	0	33	1
					S49C	20	0	25	1
Total	702	17	325	12	Total	559	7	540	20

Using equation (1) above, we obtain

$$Z = \frac{12 - 7.7038}{\sqrt{4.7647}} = 1.97;\ P = 0.024$$

Using equation (1) above, we obtain

$$Z = \frac{20 - 14.3876}{\sqrt{5.4690}} = 2.40;\ P = 0.0082$$

X chromosomes were excluded from the data discussed above because (*a*) some F_1 females died after eclosion but before they could be tested, (*b*) some tested X chromosomes remained doubtful, and (*c*) some tested F_1 females remained sterile. These excluded groups may have had an effect on the observed frequency of lethals. Two of the ways in which the frequencies of the excluded groups could have made the Basc sets appear to have a lower lethal frequency than the Bv sets are the following. First, the included groups of X chromosomes (classified as lethal or non-lethal) might have had a lower lethal frequency than any one of the excluded groups of X chromosomes. Stern *et al.*[3] showed this to be true for doubtfuls. Then, even though the true rate for Basc was the same as for Bv or higher, the frequency of

lethals for the Basc sets could have appeared the lower if the actual percentage of one or more of the excluded groups were greater for Basc than for Bv. Second, if the frequency of lethals in one or more of the Basc excluded groups were higher than that among the comparable Bv excluded group, then the Basc lethal mutation rate could have appeared lower than the Bv rate even though the percentages of the excluded groups were the same and the true frequency of lethals for Basc was the same as or higher than that for Bv. We were able to test the first possibility but not the second.

If the number of heterozygotes dying between eclosion and pairing, the number of doubtfuls, and the number of steriles gave a higher total for Basc than for Bv, they might account for the lower Basc frequency. Actually they did not, for the total loss was much higher in the Bv cultures. The mortalities, doubtfuls, and steriles out of the total numbers of F_1 females collected were 9.20, 0.47, and 2.29 per cent for Bv and 4.41, 0.34, and 2.92 per cent for Basc. The respective totals were 11.96 and 7.67 per cent, and χ^2 for the totals was 89.80, with a P value of less than 0.001. Considering only the percentage of F_1 females that died between eclosion and pairing, a preponderance of Bv females died ($\chi^2 = 163$, $P < 0.001$). When the percentages of doubtful and sterile cultures are based on the total numbers of F_1 females that were classified as nonlethal, lethal, doubtful, and sterile, the following percentages were obtained. For the Bv and the Basc experiments, 0.52 per cent (29/5,555) and 0.35 per cent (40/11,276), respectively, of the X chromosomes were designated "doubtful" ($\chi^2 = 2.55$, $P = 0.11$); and 2.52 per cent (140/5,555) and 3.05 per cent (344/11,276) were designated "sterile" ($\chi^2 = 3.75$, $P = 0.054$). The total percentages of doubtfuls and steriles were 3.04 for Bv and 3.40 for Basc ($\chi^2 = 1.54$, $P = 0.22$). Therefore, the actual ratios of doubtfuls and steriles cannot account for the low rate of the Basc tests if the first possibility discussed above is correct.

Another source of the difference in the lethal frequencies might be the different genetic constitutions of the F_2 and F_3 wild-type males of the Basc and Bv sets. Whether or not the X chromosome of a wild-type male was classified as lethal may have depended not only on the X chromosome but on the Y chromosome and (or) autosomes, at least for some lethals. Although all the wild-type males in even one culture probably would not carry the same autosomes, a difference in the over-all average of the two lines may have been effective in bringing about the observed difference in lethal frequency. The Basc F_1 offspring (from the cross, Basc female $\times$ Samarkand male) consisted of Basc/+ females and Basc males, except for the possible exceptions due to nondisjunction. However, the Bv F_1 offspring (Bv female $\times$ Samarkand male) consisted of *yvf · YL*/+ and *y sc*^{s1} *B In49 v w*^{a} *sc*^{s}/+ females and *y v f · YL* and *y sc*^{s1} *B In49 v w*^{a} *sc*^{s} males. As these B/+ females were not collected as virgins, they had time to mate with both types of male sibs before they were stored with Basc males. Fertile matings with *yvf · YL* males were rare and gave few offspring from each mating; for 39 Bv F_2 lethals, counts were recorded of classes and numbers of flies in the F_2 generation, and, in three of these, offspring of *y v f · YL* males were observed, although in the same cultures offspring from males with chromosomes containing the inversions were far more numerous. Because the F_1 females could mate with sibs before they were stored with males from the Basc stocks, wild-type males in the F_2 and F_3 generations varied in their

proportions and types of autosomes and proportions of Samarkand and Basc Y chromosomes, depending upon whether the P female was Bv or Basc. This difference was enhanced by the fact that F_1 daughters of Basc females were stored with Basc male sibs and males from the Basc stock, while F_1 daughters of Bv females were stored with males from the Basc stock only. It is conceivable that the autosomal and Y-chromosomal constitution of the Basc lines, which differs from that of the Bv lines, enables males carrying certain lethals in the X chromosome to eclose and to appear as normal males, while the same lethal in a Bv genetic background would actually be lethal according to the standards set up. If this is true, replacement of the chromosomes from the Bv stock by chromosomes from the Basc and Samarkand stocks could change a Bv lethal into a nonlethal. As far as Basc is concerned, replacement of this kind was accomplished to some extent in our experiments, for F_1 females and F_2 females (provisionally lethal-bearing) were always mated with Basc males. While a higher proportion of provisional lethals from the Bv females became nonlethal in the final check (10 of the original 42 provisional F_2 lethals) than did the lethals from the Basc females (2 of the original 26 provisional lethals), the difference is not significant ($\chi^2 = 2.87$, $P = 0.09$). However, since replacement was only incomplete, the possibility is not excluded that the different genetic constitutions of the F_2 and F_3 wild-type males could account for at least some of the difference in lethal frequency between the Bv and Basc sets. If the statistical analysis of the lethal frequency were based on the numbers of provisional lethals, the significance of the difference in frequency of recovered lethals from the two types of females would be even greater than is indicated by the final analysis. Of the 10 provisional lethals from Bv females that became nonlethal in the final check, most were declared nonlethal on the basis of the presence of one or two wild-type males in the culture. It is possible that these were actually semilethals and that by chance the wild-type males did not appear in the first check for lethals.

Other very different possibilities are that the two types of inseminated females might have exerted differentially a mutagenic action on the X chromosomes of the sperm from their mates at some time between insemination and fertilization of the eggs, that some factor in the egg, cytoplasmic or nuclear, may have been mutagenic, or that the two kinds of females may have differentially caused "healing" of lethals originated before mating. The experiments reported here cannot distinguish among these possibilities.

An indication that the condition of the egg might influence the frequency of chromosome aberrations recovered in the X chromosome of the sperm was shown in the work of Bonnier, who found, in one series of experiments, that the X-raying of virgin females before mating increased the aberration frequency of the X chromosomes from sperm irradiated in the male.[4] In the experiments by Bonnier[4] and Lüning[5] it was also shown that the rate of X-ray-induced aberrations in male gametes is increased if the sperm are irradiated in impregnated females rather than in the males, and the effect is influenced by the nature of the strain of females used. Another analogous situation is that described by Brown and Hannah in their studies on chromosome elimination in *D. melanogaster*, where frequency of gynanders resulting from the loss of a ring-X chromosome varied with the different genotypes of the female parent.[6]

Summary.—The frequency of spontaneous lethals in X chromosomes derived from wild-type *D. melanogaster* males was influenced by the type of female that had been inseminated. For five experiments, two with Basc females only and three with Bv females only, carried out at different times the over-all frequency of lethals that were recovered from Bv females was 7.8 times that recovered from the Basc females. For a special experiment in which each male mated with both a Basc and a Bv female the comparable ratio was 2.7 if the lethals in a cluster were counted individually and 3.8 if the cluster was counted as a single lethal. In the tests in which each male mated first with a Basc female and second with a Bv female the frequency of lethals that were recovered from Bv females was 1.96 times that recovered from Basc females. If the frequencies are actually equal, the probability of a difference as large or larger is 0.024. When Bv females were mated first and Basc females second, the comparable ratio was 3.58 and P was 0.0082.

Possible reasons for the difference in the lethal frequencies are discussed. No conclusions can be drawn as to the influence of the genetic backgrounds of the Basc or Bv flies. In the Basc and Bv cultures the different rates of mortality of F_1 females bearing the X chromosomes to be tested, of doubtful X chromosomes, and of sterile F_1 females do not support the hypothesis that these differences might cause the frequency of lethals recovered to be lower in the Basc than in the Bv tests.

It is possible that the "spouse effect" may be one that is mutagenic or one that results in the healing of lethals formed earlier. This effect may be exerted between insemination and fertilization or may be exerted after fertilization by some factor—nuclear or cytoplasmic—in the egg.

The authors wish to thank Dr. Curt Stern for his valuable suggestions and interest throughout the course of the experiments.

* This work was performed under the auspices of the United States Atomic Energy Commission.

† Present address: Division of Natural Science, Long Beach State College, Long Beach, California.

[1] H. J. Muller, *Yearbook Am. Phil. Soc.*, pp. 150–153, 1945.

[2] The authors wish to thank Dr. Robert Wijsman, of the Department of Statistics, University of California, Berkeley, for suggesting the method of statistical analysis.

[3] C. Stern, G. Carson, M. Kinst, E. Novitski, and D. Uphoff, *Genetics*, **37**, 413–449, 1952.

[4] Gert Bonnier, *Heredity*, **8**, 199–210, 1954.

[5] K. G. Lüning, *Heredity*, **8**, 211–223, 1954.

[6] Spencer W. Brown and Aloha Hannah, these PROCEEDINGS, **38**, 687–693, 1952.

*THE DEVELOPMENT IN VITRO OF CHIMERIC AGGREGATES OF DISSOCIATED EMBRYONIC CHICK AND MOUSE CELLS**

By A. Moscona

LABORATORY OF DEVELOPMENTAL BIOLOGY, ROCKEFELLER INSTITUTE FOR MEDICAL RESEARCH, NEW YORK, NEW YORK

Communicated by Paul Weiss, November 14, 1956

Various embryonic tissues and organ rudiments can be dissociated into suspensions of discrete, viable cells following treatment with Ca- and Mg-free saline and trypsin.[1–3] When cultivated in vitro under appropriate conditions, such cells reaggregate into compact clusters (Figs. 1–6), which subsequently re-establish tissue-like relationships and differentiate histotypically.[2, 3] These findings, originally established for chondrogenic, nephrogenic, and myogenic cells, have recently been extended to other embryonic tissues.[3–8]

If two different types of embryonic chick cells are intermingled in the same suspension, the resulting aggregates incorporate both types of cells; however, in the course of the further development of such heterotypic aggregates,† the diverse types of cells form distinct, histogenetically uniform groupings.[2] The problem of grouping of animal cells in its relation to morphogenesis was discussed in detail by Weiss[9] in reference to the concepts of "affinities"[10] and "coaptation";[11] its experimental implications were explored in the chick embryo[12] and in amphibian embryos[13,14] and larvae[15–17] and also under conditions of tissue culture.[3, 18–20] Several of these studies strongly suggested that cells of diverse lineages manifested characteristic preferences in establishing intercellular contacts and tissue contiguity. This view was further supported by the results of recent experiments on heterotypic aggregates of chick cells[2] which convincingly demonstrated a type-specific grouping of cells in the formation and development of such aggregates. These observations fell short of proof, however, due to the difficulty of identifying early embryonic chick cells when dissociated into discrete units in suspension; under these conditions, nearly all types of such cells look alike, and their identities in heterotypic mixtures are therefore not readily determined. The obvious solution to this impasse was to have cells marked in a way which would make them individually distinguishable in a mixed population. In searching for suitable "marker cells," an attempt was made to exploit the morphological differences between chick and mouse cells; mouse cell neclei are larger than chick cell nuclei and stain differently with basic stains and hematoxylin. Previous studies have shown that mouse and chick tissues can be successfully cultured in heterologous media[21, 22] and maintained simultaneously in culture without apparent incompatibility;[23–25] it has further been noticed that under such conditions the differences of size and staining properties of the cells and nuclei of the two species are retained.

Accordingly, the feasibility of obtaining composite aggregates, consisting of both chick and mouse cells, was explored. Preliminary experiments[26] demonstrated that aggregates formed in suspensions of intermingled chick and mouse cells incorporated, under appropriate conditions, cells of both species. Upon further cultivation, such heterologous aggregates developed histogenetically in accordance with the origin of their cellular components. Due to the differences in size and the

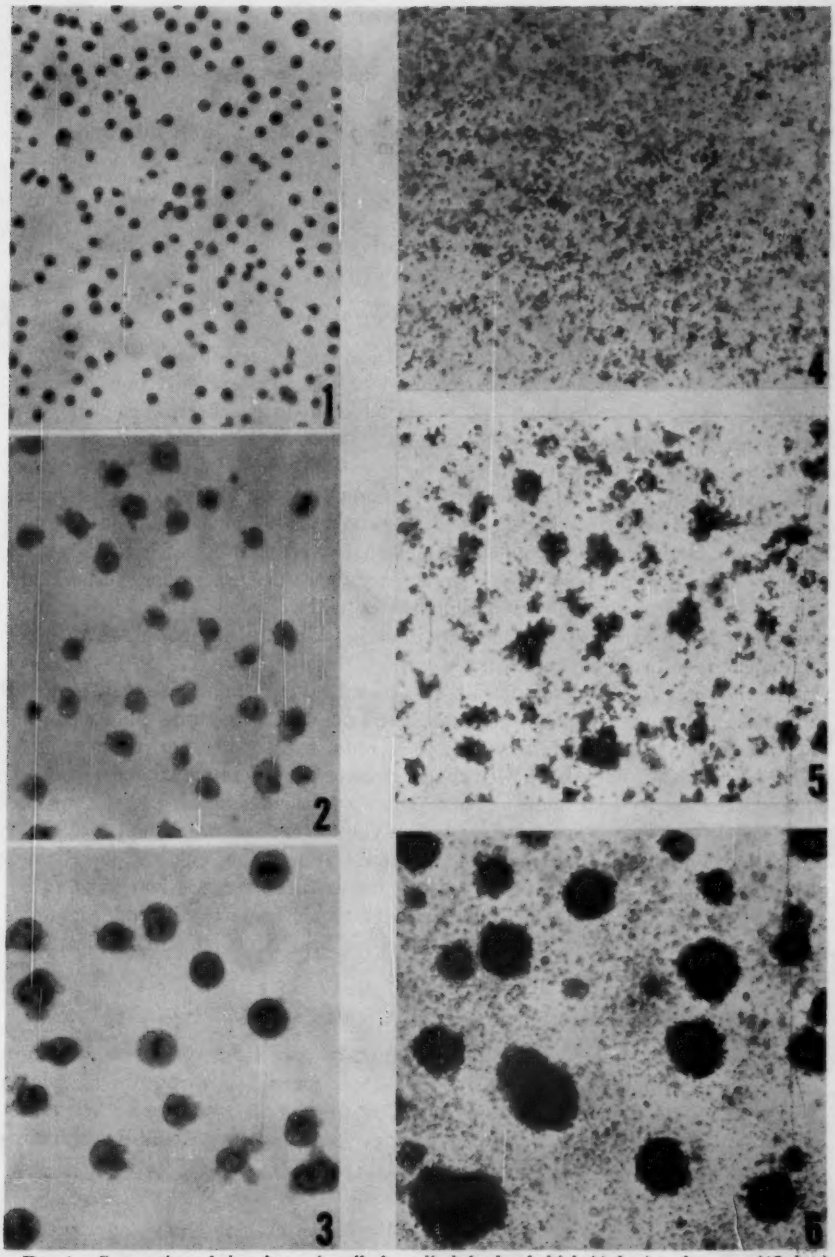


FIG. 1.—Suspension of chondrogenic cells from limb-buds of chick (4-day) and mouse (12-day) embryos. Ehrlich's hematoxylin–Biebrich's. ×160.
FIG. 2.—Suspension of chick chondrogenic cells. ×460
FIG. 3.—Suspension of mouse chondrogenic cells. ×460
FIG. 4.—Dissociated chondrogenic cells beginning to aggregate. 2-hour culture; living. ×50
FIG. 5.—A similar culture to that in Fig. 4, after 12 hours
FIG. 6.—A similar culture to that in Fig. 4, after 36 hours

staining properties of chick and mouse nuclei, the two types of cells could be easily distinguished and their precise distribution in the aggregates determined. As a further variation along this line, dissociated mouse tumor cells were introduced into suspensions of embryonic chick cells, and the structure and composition of the resulting aggregates were examined. With the aid of these differential cellular systems, various aspects of tissue reconstruction and development in cell aggregates were studied. Some of the observations bearing on the problem of cellular grouping are reported below.

MATERIALS AND METHODS

The preparation by treatment with trypsin of cell suspensions from embryonic organ rudiments and tumor tissue followed procedures described previously.[1-3] The experiments reported here were made with chondrogenic, mesonephric, and hepatic cells from chick and mouse embryos. Different age combinations of these tissues were tried, as it turned out that embryonic chick and mouse cells of diverse ages and types migrated and aggregated at different rates. This communication reports on tissues from 3- to 5-day chick embryos and from 11- to 13-day mouse embryos. The tumor tissue used was pigmented melanoma S91, maintained in a DBA/2JN strain of mice. Suspensions of cells were mixed in the desired proportions, and aliquots of the heterologous mixtures were distributed into hollow-ground (Maximow) slides with 1.0 cc. liquid culture medium in each. The medium consisted of 40 per cent chicken serum, 40 per cent embryo extract (freshly prepared from 10- to 12-day chick embryos), and 20 per cent Earl's balanced salt solution. Horse serum was sometimes added in proportions not exceeding 4 per cent of the total quantity of the medium. The culture medium was kept at room temperature for about an hour before being used. The slides with the cell suspensions were sealed and incubated at 38° C. for twenty-four hours. The medium was then changed, and the cultures maintained for an additional day or two. The aggregates which had formed by that time were then transferred to a plasma clot for further cultivation in watch glasses. After fixation in Zenker's fixative, the cultures were sectioned at 6 or 8μ, and the sections were stained briefly with Ehrlich's hematoxylin and Biebrich's scarlet, which rendered cell nuclei of the chick a light purple tint, while mouse nuclei stained a deep blue.

ISOTOPIC COMBINATIONS OF CHICK AND MOUSE CELLS

Dissociated chondrogenic cells from the limb-buds of 4-day chick embryos were thoroughly intermingled in suspension with chondrogenic cells from the limb-buds of 12-day mouse embryos (Figs. 1-3). The amount of mouse cells was about double that of chick cells. At this stage of development, the presumptive chondro-

FIG. 7.—Cartilage masses composed of interspersed chick and mouse chondrogenic cells. ×120. (FIGS. 7-19.—Stained with Ehrlich's hematoxylin-Biebrich's scarlet.)

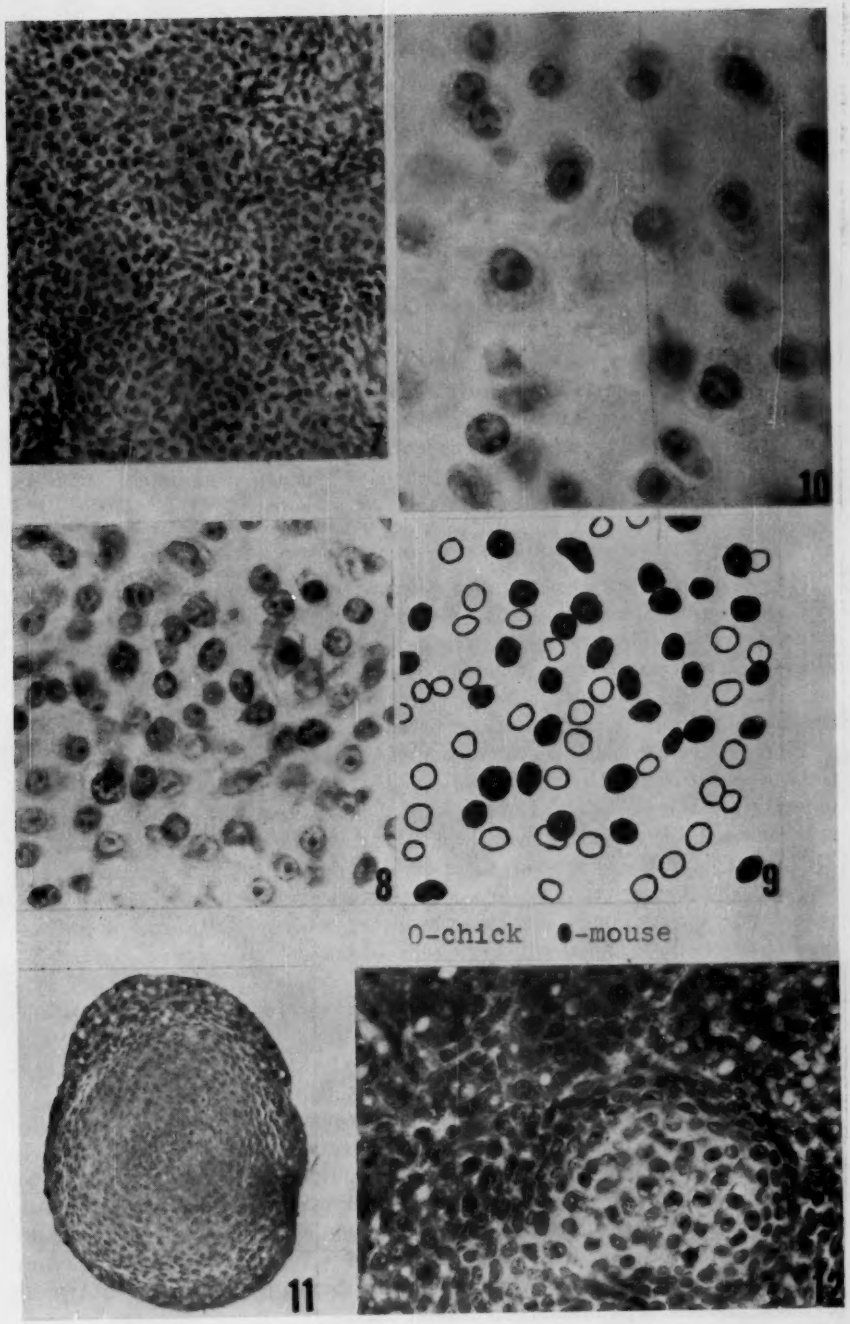
FIG. 8.—Same at ×280. Compare with Fig. 9

FIG. 9.—Outlines of nuclei of Fig. 8 to show the distribution of mouse (*circles*) and chick (*dark*) nuclei.

FIG. 10.—Full differentiated, composite cartilage, showing chick and mouse chondrocytes in a common matrix. ×980.

FIG. 11.—Aggregate of mouse liver and chick chondrogenic cells, showing a "capsule" of hepatic tissue surrounding the globule of cartilage. 4-day culture. ×80.

FIG. 12.—A 5-day culture of an aggregate of mouse hepatic and chondrogenic cells, showing the cells separated according to types. ×620.


7
10
8
9
0-chick •-mouse
11
12

blasts of the limb-bud are still in the form of stellate mesenchyme cells. The aggregates that formed in such suspensions were cultured for 6 days. Histological sections showed that they consisted of typical cartilage formed by chick and mouse cells interspersed with each other (Fig. 7). Both types of cells were intimately associated and bound by the common cartilaginous matrix into a uniform tissue fabric: the matrix surrounding a mouse cell merged quite imperceptibly with that around the chick cell next to it (Figs. 8–10). Cultures of such aggregates were maintained for periods up to one month without evidence of deterioration or incompatibility between the chick and mouse cells. Evidently the common histogenetic fabric reconstructed by the cells under these conditions was acceptable to both chick and mouse cells and suitable for their histotypical development.

An additional instance of such formative integration of interspersed chick and mouse cells was observed in combinations of liver cells. Liver tissue was obtained from 5-day chick embryos and from 13-day mouse embryos. The dissociated cells from both sources aggregated to form hepatic cords that consisted of interspersed chick and mouse cells producing glycogen or fat. In this case as well, the cells, regardless of their generic origin, reconstructed a common tissue fabric which developed in accordance with their pre-established properties.

HETEROTYPIC COMBINATIONS OF CHICK AND MOUSE CELLS

The cellular architecture of aggregates formed by cells of two diverse histogenetic types was quite different from that of isotypic cell aggregates. Mixtures of dissociated chick chondrogenic cells and mouse liver cells formed aggregates in which, after 4 days in culture, both cartilage and hepatic tissue were present. In this case, however, the two cell types had become regionally separated: the cartilage cells formed one or more central clusters, and the hepatic cells were situated around the periphery of the cartilage. In the present case, contrary to isotypic combinations, the two constituent tissues were not of mixed, chimeric composition, but each contained cells of the species that had furnished the respective cell type; that is, cartilage consisted solely of chick cells, hepatic tissue exclusively of mouse cells (Figs. 11, 12). This spatial arrangement was quite characteristic for the heterologous, as well as the homologous, combinations of these two types of cells (see also Wolff[24]).

Such type-specific grouping of cells was perhaps even more striking in combinations of mesonephric and chondrogenic cells, because of the structural characteristics of nephric tissue. In composite aggregates of 4-day chick mesonephric cells and 12-day mouse chondrogenic cells, cultured for 5 days, both kidney and cartilage cells reconstituted their recognizable tissue patterns (Figs. 13–16). The cells became consistently grouped according to type: chick chondroblasts formed areas of cartilage, mouse nephroblasts built nephric tubules. Careful examination of this material revealed no chick cells that had become chondrocytes or mouse cells that had turned into nephrocytes. Single cells that were occasionally trapped in a nonmatching environment, if they took and multiplied, developed according to their original identities. In the reversed combination of cells, namely, in aggregates of mouse mesonephric with chick chondrogenic cells, similar type-specific, separate groupings of the corresponding tissues were formed. The reconstituted nephric and chondrified showed no regular distribution within the aggregates such as was typical of combinations of hepatic and chondrogenic cells.

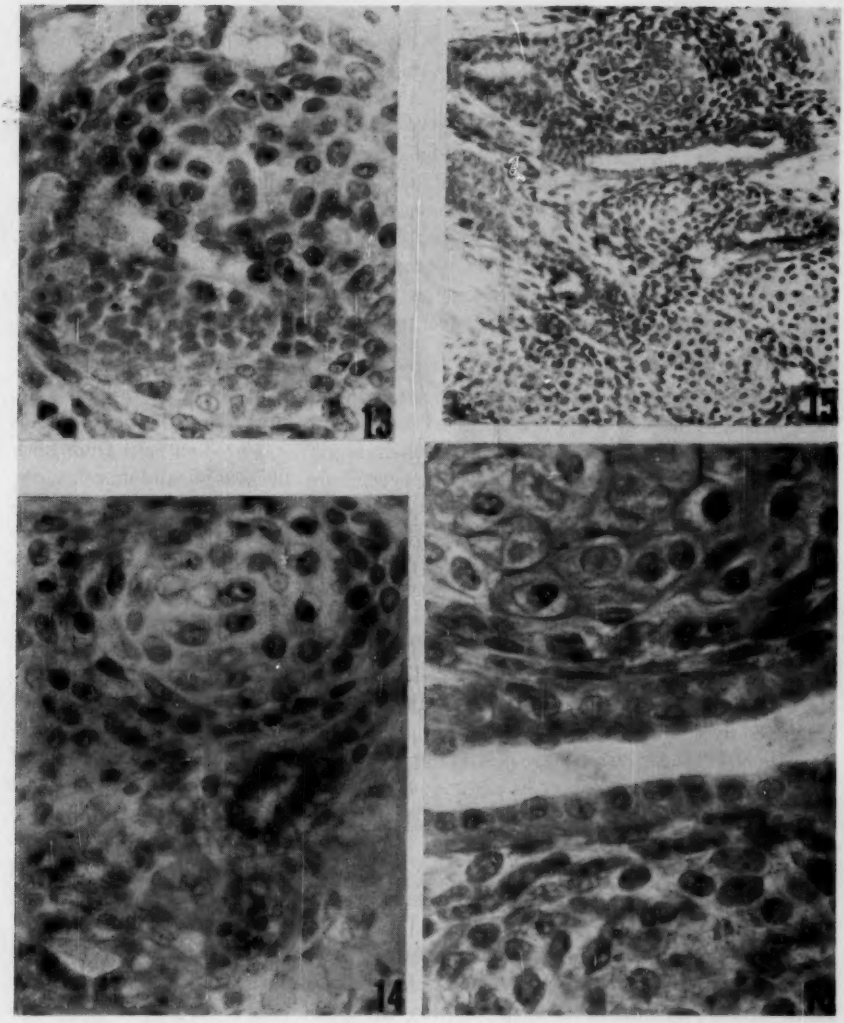


FIG. 13.—Aggregate of mouse chondrogenic and chick mesonephric cells, showing groups of cells with beginning differentiation. 3-day culture. ×250.

FIG. 14.—4-day culture of a chimeric aggregate as in Fig. 13, showing advanced histodifferentiation. ×250.

FIG. 15.—6-day culture of a chimeric aggregate as in Fig. 13, showing mosaic distribution of the cellular groupings. ×100.

FIG. 16.—Enlarged part of Fig. 15 to show the topographical proximity of the reconstituted chick and mouse tissues. ×830.

COMBINATIONS OF EMBRYONIC CHICK CELLS AND MOUSE MELANOMA CELLS

In another aspect of this study of the grouping properties of embryonic cells, observations were made on their behavior in the presence of tumor cells. Dissociated hepatic or chondrogenic cells of the chick embryo were intermingled in suspension with dissociated cells of pigmented melanoma S91 of mice (Figs. 17, 18). Embryonic and tumor cells became incorporated in common clusters, which were then further cultured for 3–5 days. Aggregates of chondrogenic chick cells and S91 cells were found to consist of a central core of cartilage surrounded by S91 cells. In older cultures, scattered melanoma cells had infiltrated into the cartilage. Aggregates of hepatic chick cells and S91 cells consisted of a central core of melanoma cells surrounded by a compact capsule of hepatic parenchyma (Fig. 19). It appears, then, that also when intermixed with tumor cells of this type, the embryonic cells clearly manifested their tendency for typewise association as well as for type-specific localization—cartilage centrally, liver peripherally.

COMMENT

The experiments reported in the foregoing demonstrated the following facts. (*a*) Chick and mouse cells, when cultured together in vitro, retained characteristics by which they could be identified as to their origins. (*b*) When cells from both species, belonging to the *same* histogenetic type, were cultured in random mixtures, they combined to form uniform chimeric tissues. (*c*) Chick and mouse cells belonging to *different* histogenetic types, however, did not readily combine but gave rise each to its discrete type-specific differentiation. Previous experiments with heterotypic cell combinations from a single species (chick embryo) had already suggested that dissociated cells tended to preserve their original type specificities and to sort out and differentiate accordingly:[2] these observations, together with the results obtained presently with cells marked clearly as to their species origin, lead to the conclusions that, under the experimental conditions explored, (1) type specificity prevailed over species specificity in guiding the association and grouping of embryonic cells of the given types of differentiation and (2) no transformation of cells of one type into another had taken place. It should be stressed at this point that these conclusions apply to cells which had evidently reached determination prior to their being dissociated, although they had not, at that time, become typically differentiated. It is thus conceivable that different results may be obtained with cells from earlier or later stages of development as well as with other types of cells, or different experimental conditions.

The problem of type-specific development in these experiments, as in the earlier ones, has two different aspects. One refers to the formation of the primary aggregates of cells, and the other to the sorting out of the cells according to kind, concurrently with or following aggregation and their subsequent differentiation. The former aspect, concerned mainly with the mechanisms of aggregation, has been only parenthetically mentioned here, and its discussion will therefore be postponed. In view of the pertinence of these problems to the observations reported, the following brief comment should be included. The formation of all types of aggregates and their histogenesis in vitro may be markedly affected by a variety of factors. Environmental changes, such as of the physical and chemical properties of the medium or the substrate, markedly influence cellular aggregation by their differential effects

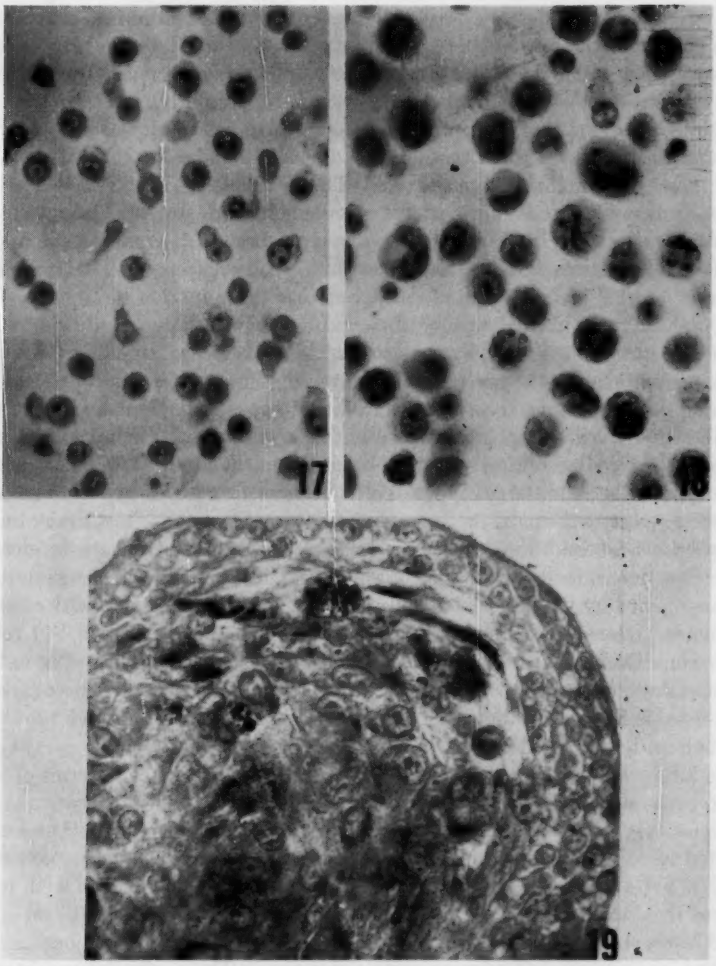


FIG. 17.—Suspension of chick liver cells. ×530
FIG. 18.—Suspension of S91 melanoma cells. ×530
FIG. 19.—Section through a composite aggregate of hepatic chick and S91 cells, cultured for 4 days and showing a "capsule" of liver tissue surrounding the cluster of melanoma cells. ×530.

on the diverse types of cells. Changes in the proportionate concentrations of different cell types intermixed in the same culture become reflected in the histological development of the ensuing aggregates.[2] As mentioned before, the rates of migration of different types of cells, as well as of cells of different generic origin, vary considerably under identical conditions. For instance, mouse mesonephric cells migrate at a slower rate than chick mesonephric cells; in cultures containing both

types of cells, this difference leads eventually to the formation of aggregates in which chick mesonephric cells predominate. Whether the different rates of migration are due to intrinsic cellular factors, to a differential response by the cells to culture conditions, or to specific activating stimuli is presently not clear.

Following formation of the primary cell aggregates, or perhaps concurrently with it, histologically identifiable tissues begin to develop, and eventually the cluster of cells becomes an organized tissue fabric. The available evidence suggests strongly that the processes of tissue formation are preceded or accompanied by a reshuffling of the aggregated cells; when more than one type of cell is incorporated in the cluster, they become sorted out to form type-specific cell groupings. The precise manner in which this occurs is still obscure, but time-lapse motion pictures, presently being undertaken in this laboratory, are expected to furnish pertinent information.

The structural differences between iso- and heterotypic cell combinations provide a striking indication of the specificities involved in cellular interactions which lead to grouping. In the tissues reconstituted from isotypic chick and mouse cells, the cells remained intermingled and interspersed in the form of cellular mosaics, without becoming segregated according to species origin. On the other hand, in aggregates of heterotypic cells the different types of cells became arranged in separate groups, so that the aggregates assumed the appearance of tissue mosaics. Thus, under the present experimental conditions, the type identities, rather than the generic identities, of the cells determined the manner of grouping. Typical grouping selectivity was also manifested by the dissociated embryonic cells when confronted with cells of the S91 tumor. However, following histogenesis of the embryonic tissue, S91 cells began, in some cases, secondarily to infiltrate between the normal cells. The nature of such manifestations, as well as the generality of such interactions, will become clearer when more is known of other combinations of dissociated normal and tumor cells and their patterns of aggregation.

The interpretation of cellular grouping in chimeric aggregates in terms of preferential, type-specific interactions between cells conforms well with observations on the tissue-specific localization of cells injected into the chick embryo[12] and into irradiated mice.[27–31] That the properties involved are effective across generic differences, not only under conditions of culture but in the organism as well, may be inferred from the successful implantations in the bone marrow of rat blood cells injected intravenously into irradiated mice.[28] In this connection, the question of the stability of chimeric cell aggregates is of interest. The successful persistence in vitro of cartilage chimeras beyond the embryonic age of their constituent cells suggested that, under such conditions, the cells, although generically alien, remained histocompatible. The response of heterologous combinations to suitable immune environments and to implantation into embryos and adults should provide additional information on the stability or the differential susceptibility of the cells under such conditions. Studies in this direction might also furnish information on the nature of histogenetic interactions between cells and the "recognition" (Weiss) effects involved, i.e., whether they function on the same basis as antibody-antigen systems[11,32,33] or whether they reflect specific properties, typical of this particular aspect of cellular behavior.

SUMMARY

1. Dissociated cells from various organ rudiments of chick and mouse embryos when intermixed in suspension cultures, readily aggregated and combined to form composite, chimeric tissues. Under suitable conditions of culture, such reconstituted tissues differentiated histotypically. This communication reports on combinations of chondrogenic, nephrogenic, and hepatogenic cells of chick and mouse embryos and S91 mouse melanoma cells.

2. In aggregates of intermixed chick and mouse cells of same type (i.e., chick and mouse chondrogenic cells) the cells reconstructed a uniform fabric which differentiated histotypically into a chimeric tissue consisting of interspersed chick and mouse cells.

3. In aggregates of intermixed chick and mouse cells of different types (i.e., chick nephrogenic and mouse chondrogenic cells) the cells became associated according to type and formed separate groupings which developed in accordance with the original histogenetic properties of the cells.

4. Due to the clear morphological differences between chick and mouse cells, it was possible precisely to identify and localize them in the chimeric aggregates. The evidence thus obtained suggested conclusively that (*a*) in the course of tissue reconstruction the dissociated embryonic cells became grouped preferentially, according to their original type identities, regardless of their generic origin, and (*b*) under the present experimental conditions no transformation of one cell type to another was observed.

The author wishes to thank Dr. Paul Weiss, head of the Laboratory of Developmental Biology, for his interest and indispensable advice throughout this study. The aid of Dr. Dorothea Bennett in some phases of this work is gratefully acknowledged.

* Research aided by grants (Paul Weiss, principal investigator) from the American Cancer Society (through the Committee on Growth, National Research Council) and the Public Health Service, National Institutes of Health.

† The following terms will be used: (1) *isotypic* and (2) *heterotypic* to designate suspensions consisting of (1) predominantly one type of cell and (2) two or more cell types; (3) *homologous* and (4) *heterologous* for cells from embryos of the same specie (3) or (4) a mixture of cells from two species (i.e., chick and mouse cells).

[1] A. Moscona, *Exptl. Cell Research*, **3**, 535, 1952.
[2] A. Moscona, *Proc. Soc. Exptl. Biol. Med.*, **92**, 410, 1956.
[3] A. Moscona and H. Moscona, *J. Anat.*, **86**, 287, 1952.
[4] P. Weiss and R. James, *Exptl. Cell Research*, Suppl. 3, p. 381, 1955.
[5] M. W. Cavanaugh, *J. Exptl. Zool.*, **128**, 573, 1955.
[6] M. W. Cavanaugh, *Exptl. Cell Research*, **9**, 42, 1955.
[7] C. Grobstein, *J. Exptl. Zool.*, **130**, 319, 1955.
[8] J. P. Trinkaus and P. W. Groves, these PROCEEDINGS, **41**, 784, 1955.
[9] P. Weiss, *Yale J. Biol. Med.*, **19**, 235, 1947.
[10] J. Holtfreter, *Arch. f. exptl. Zellforsch.*, **23**, 620, 1939.
[11] P. Weiss, *Quart. Rev. Biol.*, **25**, 177, 1950.
[12] P. Weiss and G. Andres, *J. Exptl. Zool.*, **121**, 449, 1952.
[13] J. Holtfreter, *J. Morphol.*, **80**, 25, 1947.
[14] P. L. Townes and J. Holtfreter, *J. Exptl. Zool.*, **128**, 53, 1955.
[15] J. J. Chiakulas, *J. Exptl. Zool.*, **121**, 383, 1952.
[16] F. Baltzer, *Rev. suisse Zool.*, **48**, 413, 1941.

[17] G. Andres, *Genetica*, **24**, 1, 1949.
[18] M. A. Willmer, in *Essays on Growth and Form* (Oxford: Oxford Univ. Press, 1945).
[19] V. C. Twitty and M. C. Niu, *J. Exptl. Zool.*, **108**, 405, 1948.
[20] M. Abercrombie and J. E. M. Heaysman, *Exptl. Cell Research*, **5**, 111, 1953.
[21] H. B. Fell and H. Gruneberg, *Proc. Roy. Soc. London, B*, 127, 257, 1939.
[22] H. B. Fell, *Science Progr.*, No. 162, p. 212, 1953.
[23] C. Grobstein and J. S. Youngner, *Science*, **110**, 501, 1949.
[24] E. Wolff, *Bull. Soc. Zool. France*, **79**, 357, 1954.
[25] E. Wolff and D. Bresch, *Compt. rend. Acad. sci.* (Paris), **240**, 1014, 1955.
[26] Reported at the International Congress of Developmental Biology at Brown University, Providence, Rhode Island, July, 1956.
[27] D. L. Lindsley, T. T. Odell, Jr., and F. G. Tausche, *Proc. Soc. Exptl. Biol. Med.*, **90**, 512, 1955.
[28] C. E. Ford, J. Hamerton, D. W. H. Barnes, and J. F. Loutit, *Nature*, **177**, 452, 1956.
[29] C. L. Miller, *Nature*, **178**, 142, 1956.
[30] N. A. Mitchison, *Brit. J. Exptl. Pathol.*, **37**, 239, 1956.
[31] E. S. Russel, L. J. Smith, and F. A. Lawson, *Science*, **124**, 1076, 1956.
[32] R. E. Billingham, L. Brent, and P. B. Medawar, *Nature*, **172**, 603, 1953.
[33] A. Tyler, in *Analysis of Development* (Philadelphia: W. B. Saunders Co., 1955).

INFORMATION TO CONTRIBUTORS

THE PROCEEDINGS is the official organ of the NATIONAL ACADEMY OF SCIENCES and of the NATIONAL RESEARCH COUNCIL for the publication of brief accounts of important current researches of members of the ACADEMY and of the COUNCIL and of other American investigators. The PROCEEDINGS will aim especially to secure prompt publication of original announcements of discoveries and wide circulation of the results of American research among investigators in other countries and in all branches of science.

ARTICLES should be brief. The viewpoint should be comprehensive in giving the relation of the paper to previous publications of the author or of others and in exhibiting, where practicable, the significance of the work for other branches of science. Elaborate technical details of the work and long tables of data should be avoided, but authors should be precise in making clear the new results and should give some record of the methods and data upon which they are based.

MANUSCRIPTS should be prepared with a current number of the PROCEEDINGS as a model in matters of form, and should be typewritten in duplicate with double spacing, the author retaining one copy. Illustrations should be confined to text figures of simple character, though more elaborate illustration may be allowed in special instances to authors willing to pay for their preparation and insertion. Particular attention should be given to arranging tabular matter in a simple and concise manner.

MANUSCRIPTS will be accepted only from members of the Academy, who will assume responsibility for the propriety and scientific standards of the paper and for any printing costs in excess of those allowed them by the Academy. Manuscripts should be addressed to *Proceedings of the* NATIONAL ACADEMY OF SCIENCES, Attention Miss Mary D. Alexander, University of Chicago Press, 5750 Ellis Avenue, Chicago 37, Illinois.

GALLEY PROOF will be sent, with the understanding that charges for author's corrections shall be billed to him. Author should therefore make final revisions on the typewritten manuscripts. Page proofs will not be sent.

REPRINTS should be ordered when proof is returned, on the reprint order form which will be inclosed with the proof. The order blank gives a tabular estimation of the cost of reprints; a copy will be supplied at any time upon request.

CONTENTS

		Page
GEOPHYSICS.—INTRODUCTORY REMARKS	*By Ernest H. Vestine*	1
GEOPHYSICS.—INFERENCES FROM RADIO SIGNALS FROM THE SUN AND PLANETS	*By John Firor*	2
GEOPHYSICS.—THE SOLAR HYDROMAGNETIC DYNAMO	*By E. N. Parker*	8
GEOPHYSICS.—THE TERRESTRIAL DYNAMO	*By Walter M. Elsasser*	14
GEOPHYSICS.—ON COSMIC MAGNETIC FIELDS	*By S. Chandrasekhar*	24
GEOPHYSICS.—SOLAR INFLUENCES ON COSMIC RAYS	*By S. E. Forbush*	28
GEOPHYSICS.—SOLAR ORIGIN OF CHANGES IN THE PRIMARY COSMIC RADIATION	*By A. Simpson*	42
GEOPHYSICS.—DIRECT DETECTION OF AURORAL RADIATION WITH ROCKET EQUIPMENT	*By James A. Van Allen*	57
GEOPHYSICS.—PROBLEMS IN AURORAL MORPHOLOGY	*By C. T. Elvey*	63
GEOPHYSICS.—SOLAR INFLUENCES OF GEOMAGNETISM	*By Julius Bartels*	75
GEOPHYSICS.—OBSERVATIONAL AND THEORETICAL ASPECTS OF MAGNETIC AND IONOSPHERIC STORMS	*By Ernest H. Vestine*	81
GEOPHYSICS.—VARIATIONS OF THE IONOSPHERE AND OF THE SOLAR INTENSITY WITH SUN-SPOTS	*By E. O. Hulburt*	92
GEOPHYSICS.—SOLAR RADIATION AND THE LOWER ATMOSPHERE	*By Sigmund Fritz*	95
BIOCHEMISTRY.—ARE CHLOROPLASTS SEMICONDUCTORS?	*By William Arnold and Helen K. Sherwood*	105
BIOCHEMISTRY.—STUDIES ON THE IMMUNOLOGICAL OVERLAP AMONG CERTAIN ARTHROPOD-BORNE VIRUSES. II. THE ROLE OF SEROLOGIC RELATIONSHIPS IN EXPERIMENTAL VACCINATION PROCEDURES	*By Winston H. Price*	115
BIOCHEMISTRY.—THE ORGANIZATION AND DUPLICATION OF CHROMOSOMES AS REVEALED BY AUTORADIOGRAPHIC STUDIES USING TRITIUM-LABELED THYMIDINE	*By J. Herbert Taylor, Philip S. Woods, and Walter L. Hughes*	122
BIOCHEMISTRY.—A METHOD FOR INCREASING BRAIN SEROTONIN WITHOUT INCURRING SOME OF THE PERIPHERAL EFFECTS OF THE HORMONE	*By D. W. Woolley, E. Van Winkle, and E. Shaw*	128
BOTANY.—SOME FACTORS INFLUENCING THE LONG-WAVE LIMIT OF PHOTOSYNTHESIS	*By Robert Emerson, Ruth Chalmers, and Carl Cederstrand*	133
CHEMISTRY.—THE SYMMETRY OF DENDRITIC SNOW CRYSTALS	*By Dan McLachlan, Jr.*	143
CHEMISTRY.—EXCITATIONS AND POLYMERIZATION	*By Albert Szent-Györgyi*	151
GENETICS.—MALE SEGREGATION RATIO ADVANTAGE AS A FACTOR IN MAINTAINING LETHAL ALLELES IN WILD POPULATIONS OF HOUSE MICE	*By David Bruck*	152
GENETICS.—EVIDENCE OF EVOLUTIONARY FORCES LEADING TO THE SPREAD OF LETHAL GENES IN WILD POPULATIONS OF HOUSE MICE	*By L. C. Dunn*	158
MATHEMATICS.—COMBINATORIAL PROBLEMS IN THE THEORY OF GRAPHS. IV	*By G. W. Ford and G. E. Uhlenbeck*	163
MATHEMATICS.—INDICES OF RANK AND OF SINGULARITY ON ABELIAN VARIETIES	*By R. C. Gunning*	167
MATHEMATICS.—ON DEHN'S LEMMA AND THE ASPHERICITY OF KNOTS	*By C. D. Papakyriakopoulos*	169
MATHEMATICS.—FUNCTION RINGS ON THE CIRCLE	*By J. Wermer*	173
ZOÖLOGY.—INFLUENCE OF THE TYPE OF INSEMINATED FEMALE ON THE LETHAL FREQUENCY IN THE X CHROMOSOME FROM THE MALE	*By Philip E. Hildreth and Gweneth L. Carson*	175
ZOÖLOGY.—THE DEVELOPMENT IN VITRO OF CHIMERIC AGGREGATES OF DISSOCIATED EMBRYONIC CHICK AND MOUSE CELLS	*By A. Moscona*	184